MASTERARBEIT

Acceptance of AI Virtual Assistants in the Context of Housing Modernization

vorgelegt am 17. November 2026

Kilian Hildebrandt

Erstprüferin: Prof. Dr. Larissa Putzar
Zweitprüfer: M.Sc. Thorben Ortmann

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Fakultät Informatik und Digitale Gesellschaft
Finkenau 35
22081 Hamburg

Zusammenfassung

Die energieeffiziente Modernisierung von Mietwohnungen wird durch komplexe technische und rechtliche Fragen gebremst. Virtuelle KI-Assistenten (AIVAs) auf Basis großer Sprachmodelle könnten dieses Wissen zugänglicher machen, vorausgesetzt, die Nutzer akzeptieren sie. Diese Arbeit evaluiert das UVA-UTAUT-Modell, das die „Unified Theory of Acceptance and Use of Technology“ um die wahrgenommene Usability und den gezeigten Affekt erweitert, und untersucht, ob ein kleines Sprachmodell anders wahrgenommen wird als ein größeres. Im Rahmen des IntelMOD-Projekts wurden zwei RAG-basierte AIVAs entwickelt: ein geführter Assistent und ein offener Chat, die jeweils ein Apertus-Modell mit entweder 8 oder 70 Milliarden Parametern nutzten. In einer gegenbalancierten Laborstudie ($N = 24$) nutzten die Teilnehmer beide Anwendungen, während ihre Mimik mit AFFDEX analysiert wurde, und bewerteten sie anschließend mit dem SUS und dem UTAUT-Fragebogen. Die wahrgenommene Usability hing stark und in ähnlichem Maße mit der Leistungserwartung und der Aufwandserwartung zusammen. Der geführte Assistent erhielt höhere Usability-Bewertungen als der offene Chat. Die gezeigte Valenz stand nicht in einem positiven Zusammenhang mit der Akzeptanz, und die Erregung moderierte diesen Zusammenhang nicht. Die Modellgröße hatte keinen signifikanten Einfluss auf die Bewertungen, ließ sich jedoch nicht von der Hosting-Umgebung trennen. Die Ergebnisse stützen die wahrgenommene Usability als zentralen Faktor für die Akzeptanz von AIVAs und warnen davor, die automatisierte Gesichtsausdrucksanalyse als alleinigen Prädiktor für die Akzeptanz zu verwenden.

Stichworte: Virtuelle Assistenten, Technologieakzeptanz, UTAUT, Affective Computing, Usability, Gebäudemodernisierung

Abstract

The energy-efficient modernization of rental housing in Germany is held back by complex technical and legal questions. AI virtual assistants (AIVAs) based on large language models could make this knowledge more accessible, provided that users accept them. This thesis evaluates the UVA-UTAUT, which extends the Unified Theory of Acceptance and Use of Technology with perceived usability and displayed affect, and examines whether a small language model is perceived differently from a larger one. A guided assistant and an open chat, each running an Apertus language model with either 8 or 70 billion parameters, were developed as two RAG-based AIVAs in the IntelMOD project. In a counterbalanced laboratory user study ($N = 24$), participants used both applications while their facial expressions were analyzed with AFFDEX, and then rated them with the SUS and the UTAUT questionnaires. Perceived usability was strongly and similarly related to performance expectancy and effort expectancy. The guided assistant received higher usability ratings than the open chat. Displayed valence was not positively related to acceptance, and arousal did not moderate this relationship. The language model size had no significant effect on any rating, but it was confounded with local and remote hosting. The results support perceived usability as a central factor in AIVA acceptance and caution against using automated facial expression analysis as a sole predictor of acceptance.

Keywords: AI virtual assistants, technology acceptance, UTAUT, affective computing, usability, housing modernization

Contents

List of Figures IV

List of Tables VI

Abbreviations VIII

1 Introduction 1

2 Fundamentals 4

2.1 Technology Acceptance . . . 4
2.1.1 Pre-UTAUT . . . 4
2.1.2 UTAUT . . . 12
2.1.3 Post-UTAUT . . . 16
2.2 Affective Computing . . . 17
2.2.1 Emotion Models . . . 17
2.2.2 Emotion Recognition Strategies . . . 18
2.2.3 FER in Practice . . . 20
2.2.4 Ethics and Bias . . . 20
2.3 User Experience & Usability . . . 21
2.3.1 Usability . . . 21
2.3.2 User Experience . . . 22
2.3.3 Relationship between UX and Usability . . . 23
2.3.4 Assessing Perceived Usability and UX . . . 23
2.4 Large Language Models . . . 27
2.4.1 Architecture . . . 28
2.4.2 Alignment . . . 28
2.4.3 Limitations and Retrieval-Augmented Generation . . . 28
2.4.4 Model Overview . . . 29

3 Related Work 31

3.1 Acceptance of AIVAs . . . 31
3.1.1 Affect and Technology Acceptance . . . 31
3.1.2 Usability and Technology Acceptance . . . 32

3.1.3 The UVA-UTAUT . . . 33
3.1.4 Acceptance of AI-Based Assistants . . . 34
3.2 Research on AIVAs . . . 34
3.3 Research Gaps . . . 35

4 Research Design **37**
4.1 Hypotheses . . . 37
4.2 Research Plan . . . 39

5 System Design and Implementation **41**
5.1 Context and Design Goals . . . 41
5.2 Requirements . . . 42
5.3 Design Constraints . . . 42
5.4 System Architecture . . . 45
5.5 Retrieval-Augmented Generation . . . 49
5.6 Technical Validation . . . 50

6 Methodology **53**
6.1 Study Design . . . 53
6.2 Measures . . . 54
6.3 Study Procedure . . . 55
6.4 Recruiting and Ethical Considerations . . . 58
6.5 Analysis Strategy . . . 58

7 Results **60**
7.1 Sample Description . . . 60
7.1.1 Demographic Questionnaire . . . 60
7.1.2 Application Usage . . . 60
7.2 Descriptive and Inferential Statistics . . . 62
7.2.1 SUS . . . 62
7.2.2 Valence and Arousal . . . 64
7.2.3 UTAUT . . . 65
7.3 Testing Hypotheses . . . 68
7.3.1 Usability × Technology Acceptance (H1a, H1b, H1c) . . . 68
7.3.2 Valence and Arousal × Technology Acceptance (H2a, H2b, H2c) . . . 70
7.3.3 Categorical Emotions × Technology Acceptance . . . 71
7.3.4 Application and Model Size (H3a–H4c) . . . 74
7.4 Exploratory Findings . . . 74
7.4.1 Demographics × Technology Acceptance . . . 76
7.4.2 Application Usage × Technology Acceptance . . . 77

8 Discussion **79**
8.1 Evaluating Hypotheses 79
8.2 Answering the Research Questions 83
8.3 Model Refinement 86
8.4 Implications 86
8.5 Limitations 87

9 Conclusion **89**
9.1 Summary 89
9.2 Future Work 90

Bibliography **91**

Appendix **109**

List of Figures

2.1 Theory of Reasoned Action (adapted from [26]) . . . 5
2.2 Technology Acceptance Model (adapted from [10]) . . . 5
2.3 Technology Acceptance Model 2 (adapted from [30]) . . . 6
2.4 Theory of Planned Behavior (adapted from [31]) . . . 7
2.5 Combined TAM-TPB (adapted from [33]) . . . 7
2.6 Decomposed TPB (adapted from [34]) . . . 8
2.7 Social Cognitive Theory (adapted from [35]) . . . 9
2.8 Social Cognitive Theory in IS (adapted from [36]) . . . 9
2.9 Model of PC Utilization (adapted from [38]) . . . 10
2.10 Five perceived attributes of innovations (adapted from [39]) . . . 11
2.11 Integration of IDT and TRA (adapted from [41]) . . . 11
2.12 Motivational Model (adapted from [42]) . . . 12
2.13 Genealogy of the models unified by the UTAUT . . . 14
2.14 Unified Theory of Acceptance and Use of Technology (adapted from [11]) . . . 16
2.15 Plutchik's wheel of emotion [55] . . . 18
2.16 Russell's circumplex model of emotions . . . 19
2.17 The CUE model (components of user experience) [86] . . . 24
2.18 RAG pipeline with indexing, retrieval, and generation . . . 29

3.1 Usability-Valence-Arousal UTAUT (adapted from [20]) . . . 33

4.1 Task sequences of the first four participants (TP1–TP4) . . . 40

5.1 Overview of Application A: Guided Assistant (German user interface) . . . 43
5.2 Overview of Application B: Basic Chat (German user interface) . . . 43
5.3 Architecture of Application A: Guided Assistant . . . 46
5.4 Architecture of Application B: Basic Chat . . . 47
5.5 Example process of retrieval for a technical question using the RAG component 51

6.1 Overview of the iMotions software study view . . . 55
6.2 Overview of the iMotions study builder . . . 56
6.3 Order of actions for the study . . . 57

7.1 Overview of application usage . . . 61

7.2 SUS scores by application and model size 63
7.3 Displayed valence and arousal by application and model size 65
7.4 Mean SUS and UTAUT scores by application and model size 66
7.5 Relationship between perceived usability and hypotheses H1a and H1b . . . 69
7.6 Moderated regression of PE, EE, and SI on displayed valence 72
7.7 Spearman correlations between demographic characteristics and the SUS and UTAUT ratings 76
7.8 Spearman correlations between usage metrics and the SUS and UTAUT ratings 77

List of Tables

2.1 Foundation models and their constructs . . . 13
2.2 The Sauro-Lewis curved grading scale (CGS) [95] . . . 25
2.3 Perceived usability ratings of everyday products [80], with grades from the curved grading scale (Table 2.2) . . . 26
2.4 Correlation of different usability measures with the SUS . . . 27

5.1 Comparison of the two applications . . . 42
5.2 Functional and non-functional requirements and their implementation . . . 44
5.3 Technology stack of the prototype . . . 48
5.4 Configuration of the small and medium language models . . . 48
5.5 Composition of the RAG knowledge base . . . 50
5.6 Comparison of the 8B and 70B models in the technical pilot . . . 52

6.1 Task sequences and assigned participants . . . 54

7.1 Usage by application × language model, M (SD) . . . 61
7.2 Likelihood-ratio tests for usage ($\chi^2(1)$ and Holm-corrected p) . . . 62
7.3 SUS scores by application × language model . . . 63
7.4 Likelihood-ratio tests for the SUS ($\chi^2(1)$ and Holm-corrected p) . . . 64
7.5 Displayed valence and arousal by application × language model, M (SD) . . . 64
7.6 Likelihood-ratio tests for displayed valence and arousal ($\chi^2(1)$ and Holm-corrected p) . . . 64
7.7 Internal consistency (Cronbach's α) . . . 65
7.8 UTAUT scores by application × language model, M (SD) . . . 66
7.9 Likelihood-ratio tests for the UTAUT constructs ($\chi^2(1)$ and Holm-corrected p) 67
7.10 Regression of performance expectancy (H1a) and effort expectancy (H1b) on SUS . . . 68
7.11 Comparison of the correlations of perceived usability with effort expectancy and performance expectancy (H1c) . . . 70
7.12 Hierarchical moderated regression of PE (H2a), EE (H2b), and SI (H2c) on displayed valence and arousal . . . 71
7.13 Displayed basic emotions by application, M (SD) and Mdn . . . 72

7.14 Spearman correlations between the displayed basic emotions and the UTAUT constructs 73
7.15 Prediction of the UTAUT constructs by all seven basic emotions 74
7.16 Tests of the differences between the applications (H3) and the model sizes (H4) 75
7.17 Spearman correlations between demographics and the SUS / UTAUT ratings 77
7.18 Spearman correlations between usage metrics and the SUS / UTAUT ratings 78

8.1 Overview of the hypotheses and their results 84

Abbreviations

Abbreviation	Meaning
AC	Affective Computing
AI	Artificial Intelligence
AIVA	AI Virtual Assistant
API	Application Programming Interface
AU	Action Unit
BGB	Bürgerliches Gesetzbuch (German Civil Code)
BGH	Bundesgerichtshof (Federal Court of Justice)
BI	Behavioral Intention
CAT	Consumer Acceptance of Technology
CSE	Computer Self-Efficacy
CSUQ	Computer System Usability Questionnaire
C-TAM-TPB	Combined TAM and TPB
CUE	Components of User Experience
CUQ	Chatbot Usability Questionnaire
DTPB	Decomposed Theory of Planned Behavior
EDA	Electrodermal Activity
EE	Effort Expectancy
EEG	Electroencephalography
FACS	Facial Action Coding System
FC	Facilitating Conditions
FER	Facial Expression Recognition
FK	Funktionales Kostensplitting
GDPR	General Data Protection Regulation
HCI	Human-Computer Interaction

Abbreviation	Meaning
IDT	Innovation Diffusion Theory
IMP	Intelligent Modernization Platform
IQR	Interquartile Range
IS	Information Systems
IT	Information Technology
LLM	Large Language Model
MM	Motivational Model
MPCU	Model of PC Utilization
NPS	Net Promoter Score
OLS	Ordinary Least Squares
PBC	Perceived Behavioral Control
PCI	Perceived Characteristics of Innovating
PE	Performance Expectancy
PEOU	Perceived Ease of Use
PU	Perceived Usefulness
RAG	Retrieval-Augmented Generation
SAM	Self-Assessment Manikin
SCT	Social Cognitive Theory
SI	Social Influence
SLM	Small Language Model
SUS	System Usability Scale
TAM	Technology Acceptance Model
TIB	Theory of Interpersonal Behavior
TPB	Theory of Planned Behavior
TRA	Theory of Reasoned Action
UEQ	User Experience Questionnaire
UMUX	Usability Metric for User Experience
UTAUT	Unified Theory of Acceptance and Use of Technology
UVA-UTAUT	Usability-Valence-Arousal UTAUT
UX	User Experience

1 Introduction

In the face of global warming, Germany has committed to reducing its greenhouse gas emissions by at least 65 % by 2030 compared with 1990 and to reaching net greenhouse gas neutrality by 2045 [1]. The building sector is one of the main contributors, as direct emissions from heating in households and commerce amounted to 103.4 Mt CO_2-eq. in 2025, about 16 % of national emissions, and rose by 3.4 Mt compared with the previous year [2]. This exceeds the sector's annual budget of 90.1 Mt [2], and current projections expect the sector to overshoot its cumulative 2021–2030 budget by about 110 Mt [2]. An industry association even attributes around 30 % of Germany's CO_2 emissions to the building sector, presumably because it also counts the electricity and district heat used in buildings [3]. Therefore, it is crucial to conduct energy-efficient modernization of the existing building stock. Two-thirds of the roughly 19.5 million residential buildings in Germany are categorized as part of the efficiency classes D to H, with classes G and H alone causing 50 % of the sector's energy consumption [3]. Yet, the energy-related renovation rate of residential buildings has fallen for three consecutive years from 0.88 % in 2022 to a record low of 0.67 % in 2025, while a renovation rate of about 2 % would be required to meet the climate targets [3]. The reasons for this regression are not only financial, as energy-efficient modernization is a complex process that requires expert knowledge about building technology, funding programs, and legal obligations. Homeowners often name process complexity, the effort of searching for information, and the lack of reliable expert advice as central barriers [4]. This process is further complicated in rental housing situations where tenancy laws apply. Measures that save energy are legally defined as modernization [5], which is why landlords may pass 8 % of the costs per year on to their tenants as a rent increase [6]. At the same time, landlords are obligated to keep the buildings in a livable state by conducting regular maintenance. Separating modernization from maintenance costs is difficult, as no established, standardized way of determining the distribution exists, often resulting in lawsuits between tenants and landlords [7].

The research projects KIMM [7] and IntelMOD [8] address this issue by building on the method of the Funktionales Kostensplitting (FK) [9]. David [9] developed the FK method with the goal of creating an understandable and reliable system to separate the modernization from the maintenance costs by assigning different functional values to building components. Although the FK method is an improvement, it is still not easy to understand for laypeople.

KIMM and IntelMOD aim to solve this by developing an Artificial Intelligence (AI)-supported online platform that advises tenants, landlords, and owners on energy-related modernization [7], [8]. A central component of this platform is an AI Virtual Assistant (AIVA) based on a Large Language Model (LLM) that helps lay users understand the FK method and the legal framework. Within this thesis, two such AIVAs are developed. The *Guided Assistant* guides users through the FK process in a fixed sequence of dialogue phases, while the *Basic Chat* aids the user in answering questions while they fill in the forms themselves (Chapter 5).

As these assistants can only fulfill their purpose if they are accepted by the users, it is important to assess their acceptance properly. Technology acceptance research offers well-established models that help to understand how technology is perceived, with the Technology Acceptance Model (TAM) [10] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [11] being the most popular ones. However, the UTAUT and its predecessors were developed to explain the adoption of conventional Information Systems (IS) in organizational settings, and there is growing evidence that they are of limited use for AI-based systems [12], [13], [14]. AI tools are often judged by the quality of the interaction itself and offer less insight into their own workings [15]. Technology acceptance models are predominantly cognitive and leave out the user's emotions during use [16], even though negative emotions play an important role in the reasons why users abandon their conversations with chatbots [17]. Additionally, the relationship between perceived usability, one of the most frequently measured constructs in Human-Computer Interaction (HCI), and technology acceptance has rarely been studied [18], [19]. To address both of these gaps, Hildebrandt et al. [20] proposed the Usability-Valence-Arousal UTAUT (UVA-UTAUT), which extends the UTAUT with perceived usability and with displayed valence and arousal as predictors of the core acceptance constructs.

Another practical question for the deployment of AIVAs is how large the underlying LLM has to be in order to provide the necessary functions. Larger models generally perform better on benchmarks [21], but they require more energy per request [22], [23] and usually cannot be hosted locally. Local hosting also matters for data protection, as the General Data Protection Regulation (GDPR) [24] restricts the transfer of personal and sensitive data to third parties. Understanding whether users actually perceive a difference between a small local model and a substantially larger one, and assessing where the potential cutoff lies, is relevant both for HCI and for sustainable system design.

This thesis therefore investigates the acceptance of AIVAs in the building modernization context and addresses the following research questions:

RQ1: How does the perceived usability of an AIVA relate to its acceptance?

RQ2: How can the affect that users display during the interaction with an AIVA contribute to explaining its acceptance?

RQ3: Do perceived usability, displayed affect, and acceptance differ between a guided and an open AIVA conversation design?

RQ4: Do perceived usability, displayed affect, and acceptance differ between an AIVA based on a small (8B) and on a medium-sized (70B) language model?

RQ1 and RQ2 examine the two extensions of the UVA-UTAUT, whereas RQ3 and RQ4 address design decisions for the practical deployment of AIVAs. Each research question is operationalized by a group of hypotheses (H1–H4), which are derived in Chapter 4, tested in Chapter 7, and answered in Section 8.2.

The contribution of this thesis is threefold. First, it extends the pilot validation of the UVA-UTAUT [25] with a counterbalanced user study that combines standardized questionnaires with automated facial expression analysis across two applications and two language model sizes. Second, it presents the design and implementation of two LLM-based AIVAs for the domain of energy-efficient modernization in rental housing, including a Retrieval-Augmented Generation (RAG) pipeline over legal and technical sources. Third, it compares a locally deployable small LLM with a substantially larger one with regard to perceived usability, displayed affect, and acceptance.

This thesis is structured as follows:

- Chapter 2 introduces the theoretical foundations of technology acceptance, affective computing, user experience and usability, and large language models.
- Chapter 3 reviews research on the acceptance of AIVAs and derives the research gaps.
- Chapter 4 derives the hypotheses for the research questions and presents the research plan.
- Chapter 5 describes the design and implementation of the two AIVAs.
- Chapter 6 describes the user study, whose results are reported in Chapter 7.
- Chapter 8 evaluates the hypotheses, answers the research questions, and discusses the limitations of the study.
- Chapter 9 summarizes the thesis and outlines future work.

2 Fundamentals

This chapter aims to provide a thorough theoretical foundation of technology acceptance research (Section 2.1), affective computing (Section 2.2), user experience research (Section 2.3), and large language models (Section 2.4) required for this thesis.

2.1 Technology Acceptance

Technology acceptance research investigates why individuals adopt or reject new Information Technology (IT).

This section provides an overview of dominant acceptance models prior to the introduction of the UTAUT (Section 2.1.1), a thorough review of the UTAUT (Section 2.1.2), as well as an overview of acceptance models following the UTAUT (Section 2.1.3).

2.1.1 Pre-UTAUT

Since the 1980s, a large number of models have been proposed, most of which explain acceptance through the beliefs and intentions of the individual user [11]. Following Venkatesh et al. [11], these models can be traced back to five theoretical roots, namely the Theory of Reasoned Action, Social Cognitive Theory, the Theory of Interpersonal Behavior, Innovation Diffusion Theory, and Motivation Theory. Figure 2.13 shows the complete model lineage.

Theory of Reasoned Action The Theory of Reasoned Action (TRA) originates in social psychology and is one of the most fundamental theories of human behavior [26]. It assumes that a person's behavior is determined by their behavioral intention, which is a function of their attitude toward the behavior, i.e., "an individual's positive or negative feelings (evaluative affect) about performing the target behavior" [26], and their subjective norm, i.e., "the person's perception that most people who are important to him think he should or should not perform the behavior in question" [26] (Figure 2.1). A meta-analysis of 87 studies confirmed the strong predictive power of the model, with a mean correlation of $r = 0.66$ between attitude and subjective norm and intention, as well as $r = 0.53$ between intention

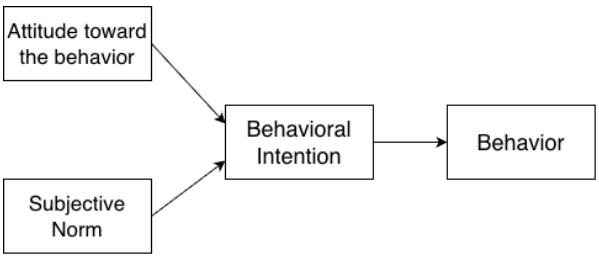

Figure 2.1: Theory of Reasoned Action (adapted from [26])

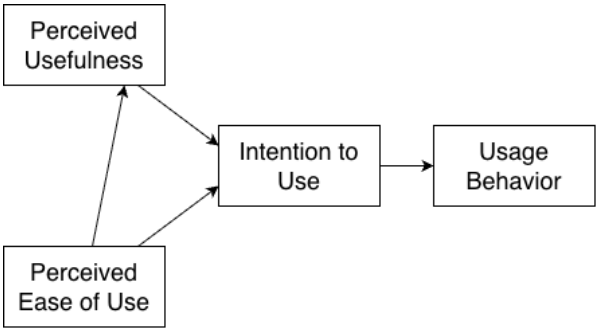

Figure 2.2: Technology Acceptance Model (adapted from [10])

and behavior [27]. However, the TRA only applies to behaviors that are "under a person's volitional control" [27]. This boundary motivated the development of the Theory of Planned Behavior.

Technology Acceptance Model Davis [10] adapted the TRA to the IT context and replaced the general beliefs with two technology-specific ones (Figure 2.2). It consists of Perceived Usefulness (PU), "the degree to which a person believes that using a particular system would enhance his or her job performance", and Perceived Ease of Use (PEOU), "the degree to which a person believes that using a particular system would be free of effort" [10]. In two validation studies, PU was much more strongly related to usage than PEOU, and the effect of PEOU on usage was largely mediated by PU [10], which led Davis [10] to conclude that "no amount of ease of use can compensate for a system that does not perform a useful function". In a direct comparison with the TRA, the TAM explained 47–51 % of the variance in intention compared with 26–32 % for the TRA [28]. Across later studies, TAM consistently explains about 40 % of the variance in intention and usage [29].

The TAM2 [30] extends the TAM by adding determinants of perceived usefulness, including subjective norm, image, job relevance, output quality, and result demonstrability (Figure 2.3) [29]. Subjective norm is reintroduced as a direct determinant of intention in mandatory settings, and its effect weakens with increasing experience [29].

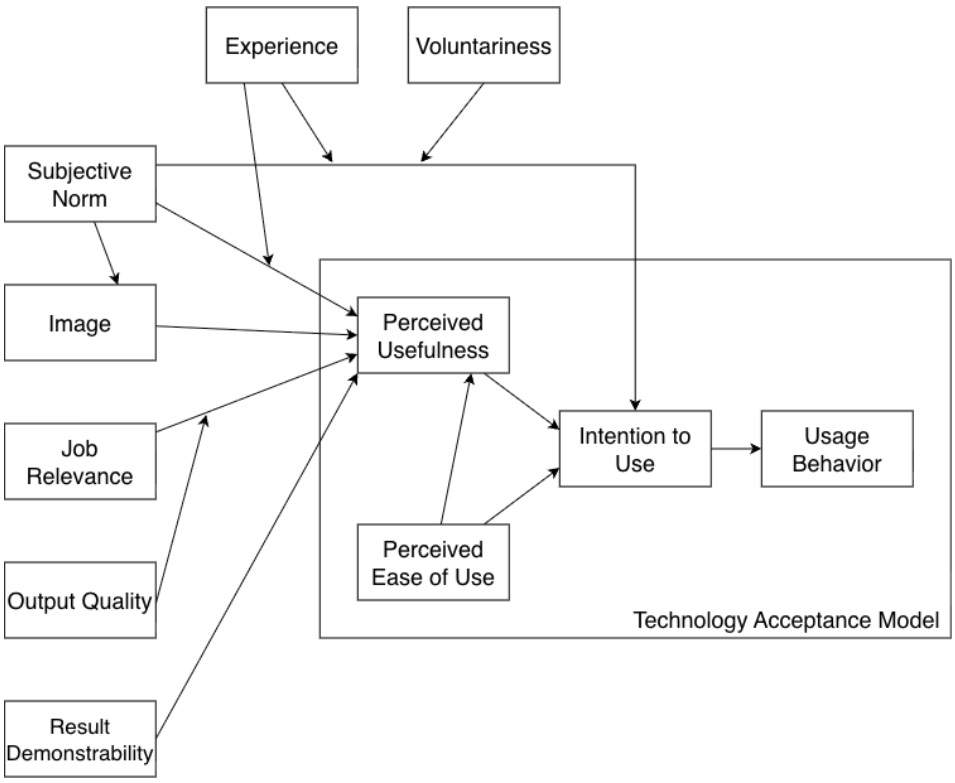

Figure 2.3: Technology Acceptance Model 2 (adapted from [30])

Theory of Planned Behavior Ajzen [31] extended the TRA with Perceived Behavioral Control (PBC) to account for behaviors that are not fully under volitional control (Figure 2.4). PBC is defined as "people's perception of the ease or difficulty of performing the behavior of interest" and affects both intention and behavior directly [31]. In the studies reviewed by Ajzen [31], attitude, subjective norm, and PBC predicted intention with a mean correlation of $r = 0.71$, and intention together with PBC predicted behavior with $r = 0.51$. Ajzen [31] describes the theory as "open to the inclusion of additional predictors" if they explain additional variance. Applied to IT, both the TAM and the Theory of Planned Behavior (TPB) predicted intention well, with TAM explaining slightly more variance [32].

Taylor and Todd [33] combined the TAM and TPB by adding subjective norm and perceived behavioral control from the TPB to the TAM (Figure 2.5). The Combined TAM and TPB (C-TAM-TPB) predicted usage for both experienced and inexperienced users, but the effect differed with experience, as perceived usefulness was a stronger predictor of intention for inexperienced users, while the link between intention and behavior was stronger for experienced users [33].

In the Decomposed Theory of Planned Behavior (DTPB), Taylor and Todd [34] took the TPB constructs and broke them down into belief structures. These include attitude into perceived usefulness, ease of use and compatibility, subjective norm into peer and superior influence, and

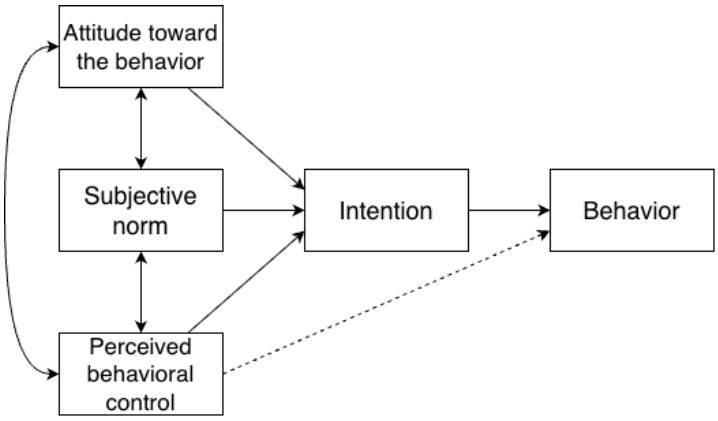

Figure 2.4: Theory of Planned Behavior (adapted from [31])

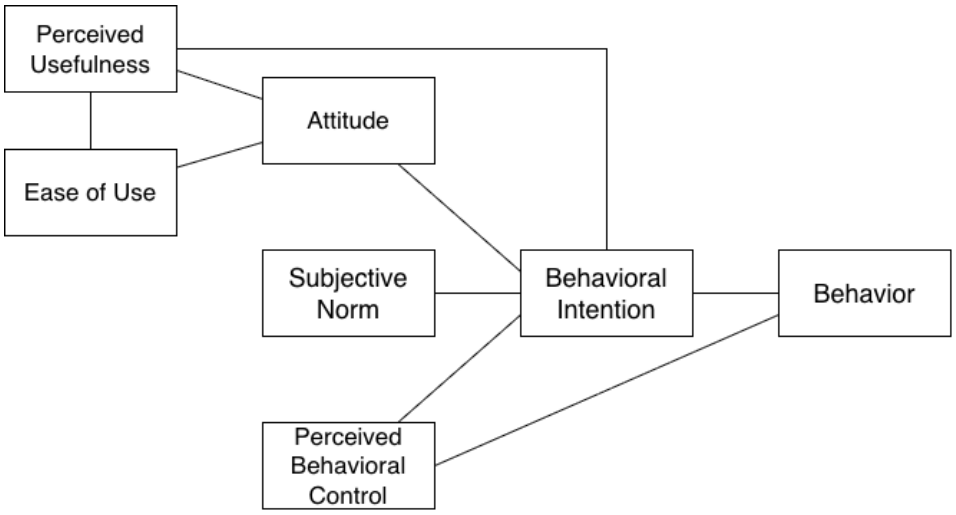

Figure 2.5: Combined TAM-TPB (adapted from [33])

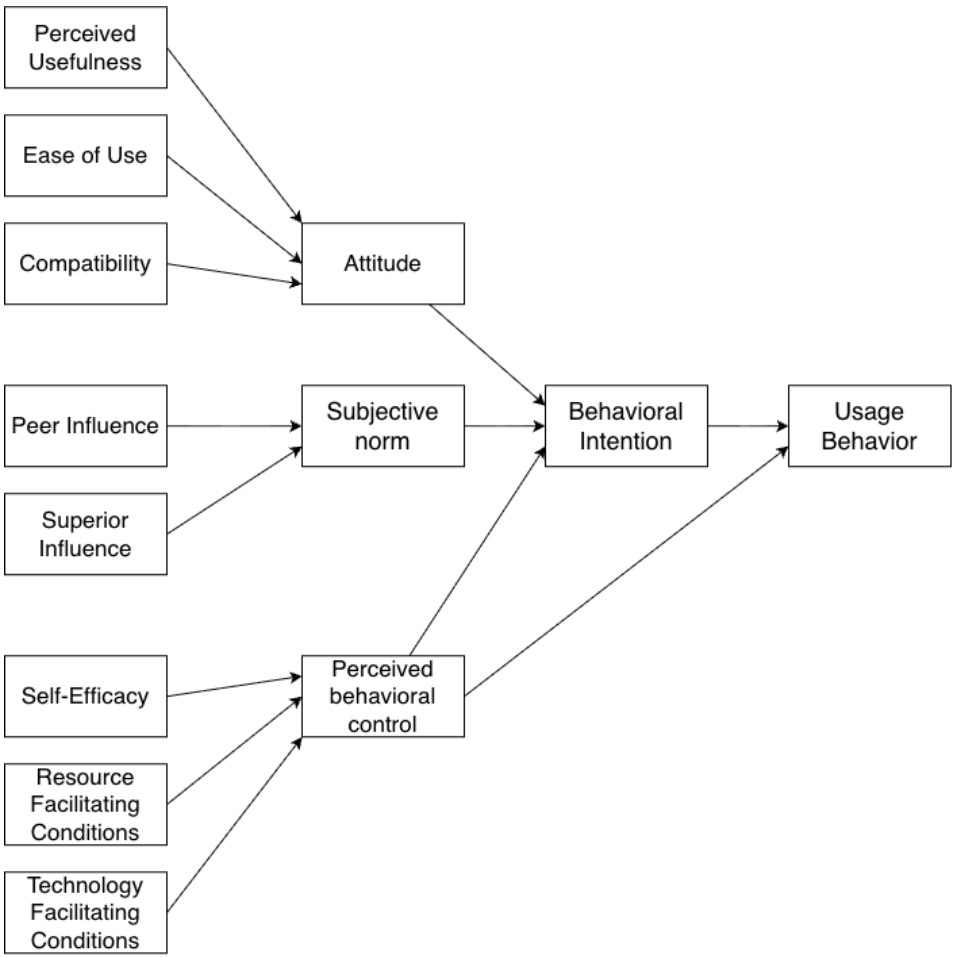

Figure 2.6: Decomposed TPB (adapted from [34])

PBC into self-efficacy, resource facilitating conditions, and technology facilitating conditions as shown in Figure 2.6. The DTPB explained slightly more variance in intention than the TAM (60 % vs. 52 %), but at the cost of seven additional constructs, gaining only two percentage points in explained behavior [34].

Social Cognitive Theory Social Cognitive Theory (SCT) explains human functioning through a triad based on behavior, cognitive and other personal factors, and environmental influences, all operating as determinants of each other (Figure 2.7) [35]. A central personal factor is self-efficacy, defined as "people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances" [35], affecting the choice of activities, effort, persistence, and emotional reactions [35]. Compeau and Higgins [36] transferred the SCT to the field of IT to assess computer use (Figure 2.8). In their model, Computer Self-Efficacy (CSE), defined as "a judgment of one's capability to use a computer" [36], influences performance- and personal-related outcome expectations, affect,

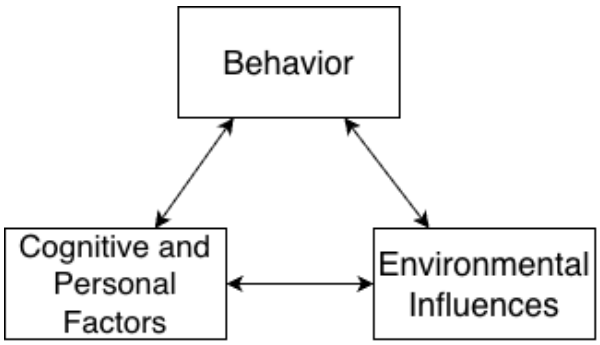

Figure 2.7: Social Cognitive Theory (adapted from [35])

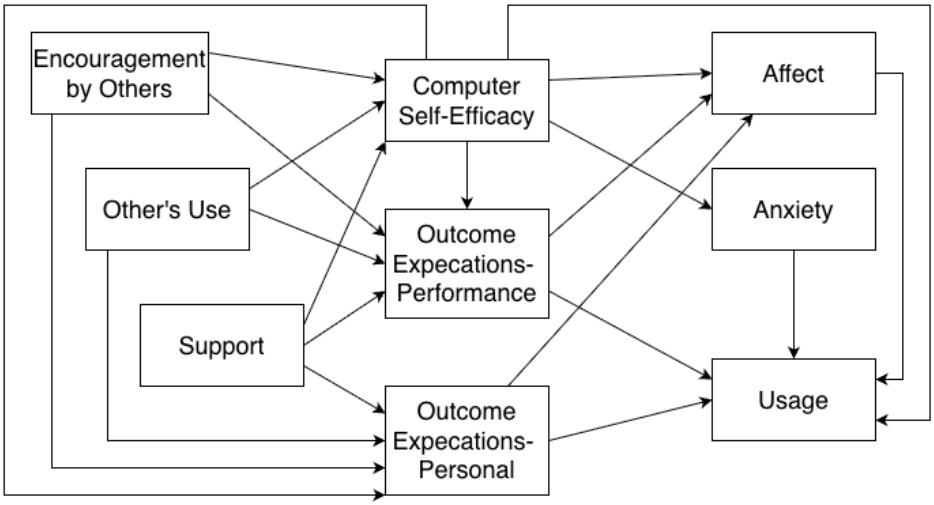

Figure 2.8: Social Cognitive Theory in IS (adapted from [36])

anxiety, and usage. CSE had a larger total effect on use than outcome expectations, and it strongly reduced computer anxiety [36].

Theory of Interpersonal Behavior The Theory of Interpersonal Behavior (TIB) by Triandis [37] served as an alternative to the TRA in the Model of PC Utilization (MPCU) by Thompson et al. [38] (Figure 2.9). The MPCU explains voluntary PC use through social factors, affect, three perceived consequences (complexity, job fit, and long-term consequences), and facilitating conditions. Job fit was the strongest predictor of utilization, while affect and facilitating conditions had no significant effect [38]. Thompson et al. [38] attribute the missing effect of affect either to PCs being perceived as tools that do not evoke strong emotions or to their model linking affect directly to use rather than through intention.

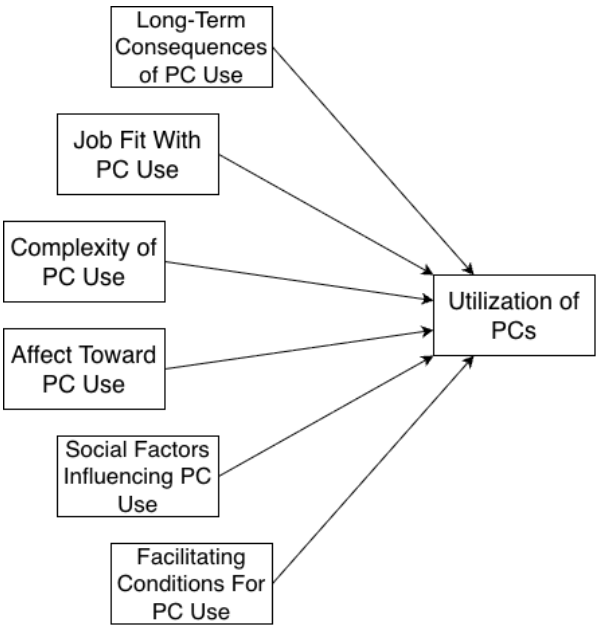

Figure 2.9: Model of PC Utilization (adapted from [38])

Innovation Diffusion Theory Innovation Diffusion Theory (IDT) describes diffusion as "the process by which an innovation is communicated through certain channels over time among the members of a social system" [39], with innovation being anything "perceived as new by an individual or other unit of adoption" [39]. Rogers [39] identifies relative advantage, compatibility, complexity, trialability, and observability as the five perceived attributes of innovations that explain 49–87 % of the variance in the rate of adoption (Figure 2.10). Moore and Benbasat [40] adapted these attributes to IT as the Perceived Characteristics of Innovating (PCI) by splitting observability into result demonstrability and visibility and adding image and voluntariness of use. Integrated into the TRA, these perceptions together with subjective norm and voluntariness explain 65 % of the variance in usage (Figure 2.11) [41].

Motivation Theory Motivation theory distinguishes between extrinsic motivation, performing an activity because it is "instrumental in achieving valued outcomes that are distinct from the activity itself", and intrinsic motivation, performing it "for no apparent reinforcement other than the process of performing the activity per se" [42]. Deci and Ryan [43] also note the innate needs for competence and self-determination. Davis et al. [42] applied this distinction to computer use, with perceived usefulness representing extrinsic and perceived enjoyment representing intrinsic motivation (Figure 2.12). Usefulness was the dominant predictor of intention, but enjoyment contributed significantly beyond it, while both also interacted positively with each other [42].

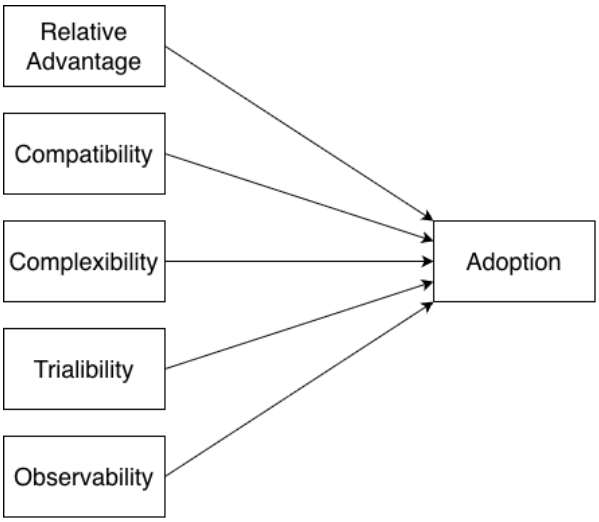

Figure 2.10: Five perceived attributes of innovations (adapted from [39])

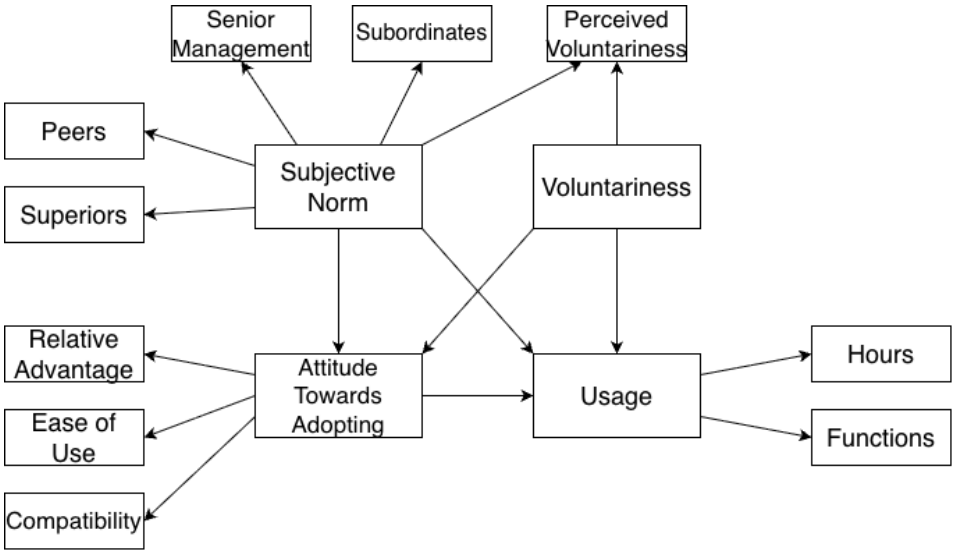

Figure 2.11: Integration of IDT and TRA (adapted from [41])

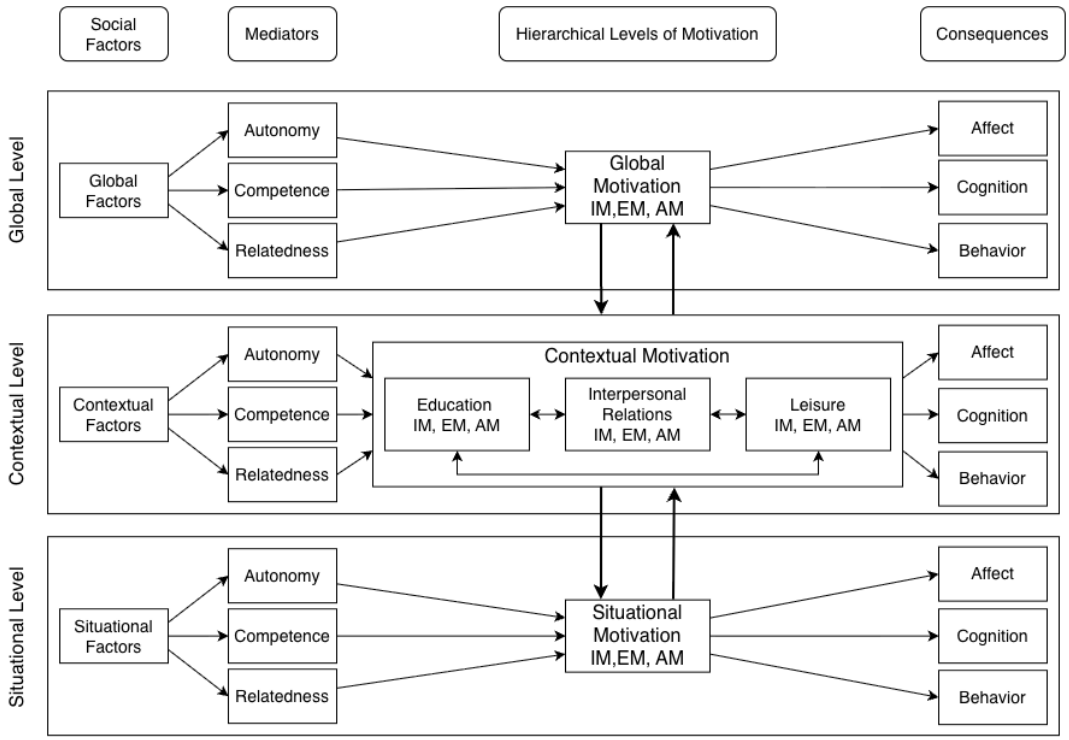

Figure 2.12: Motivational Model (adapted from [42])

2.1.2 UTAUT

The UTAUT introduced by Venkatesh et al. [11] in 2003 consolidated technology acceptance research by building a unified theoretical model on the basis of eight established models presented in Section 2.1.1 (Table 2.1 and Figure 2.13). After selecting the eight foundation models, they were compared to each other, using conceptual and empirical similarities to formulate the UTAUT. Finally, the UTAUT was tested and then cross-validated using additional collected data. The eight models used explain between 17 and 53 % of the variance in behavioral intention. The newly constructed UTAUT explains 69 % of the variance in intention in the initial test and 70 % in cross-validation, and 48 % of the variance in usage [11].

Determinants and Moderators of Technology Acceptance Venkatesh et al. [11] identified seven constructs from the foundation models as potential direct determinants of intention or usage (Figure 2.14). Four of them are theorized to play a significant role:

Performance Expectancy (PE) "the degree to which an individual believes that using the system will help him or her to attain gains in job performance" [11]. Based on perceived usefulness, extrinsic motivation, job-fit, relative advantage, and outcome expectations.

Table 2.1: Foundation models and their constructs

Model	Core Constructs
TRA	Attitude Toward Behavior Subjective Norm
TAM	Perceived Usefulness Perceived Ease of Use Subjective Norm (TAM2 only)
Motivational Model (MM)	Extrinsic Motivation Intrinsic Motivation
TPB	Attitude Toward Behavior Subjective Norm Perceived Behavioral Control
C-TAM-TPB	Attitude Toward Behavior Subjective Norm Perceived Behavioral Control Perceived Usefulness
MPCU	Job-Fit Complexity Long-term Consequences Affect Toward Use Social Factors Facilitating Conditions
IDT	Relative Advantage Ease of Use Image Visibility Compatibility Results Demonstrability Voluntariness of Use
SCT	Outcome Expectations Performance Outcome Expectations Personal Self-efficacy Affect Anxiety

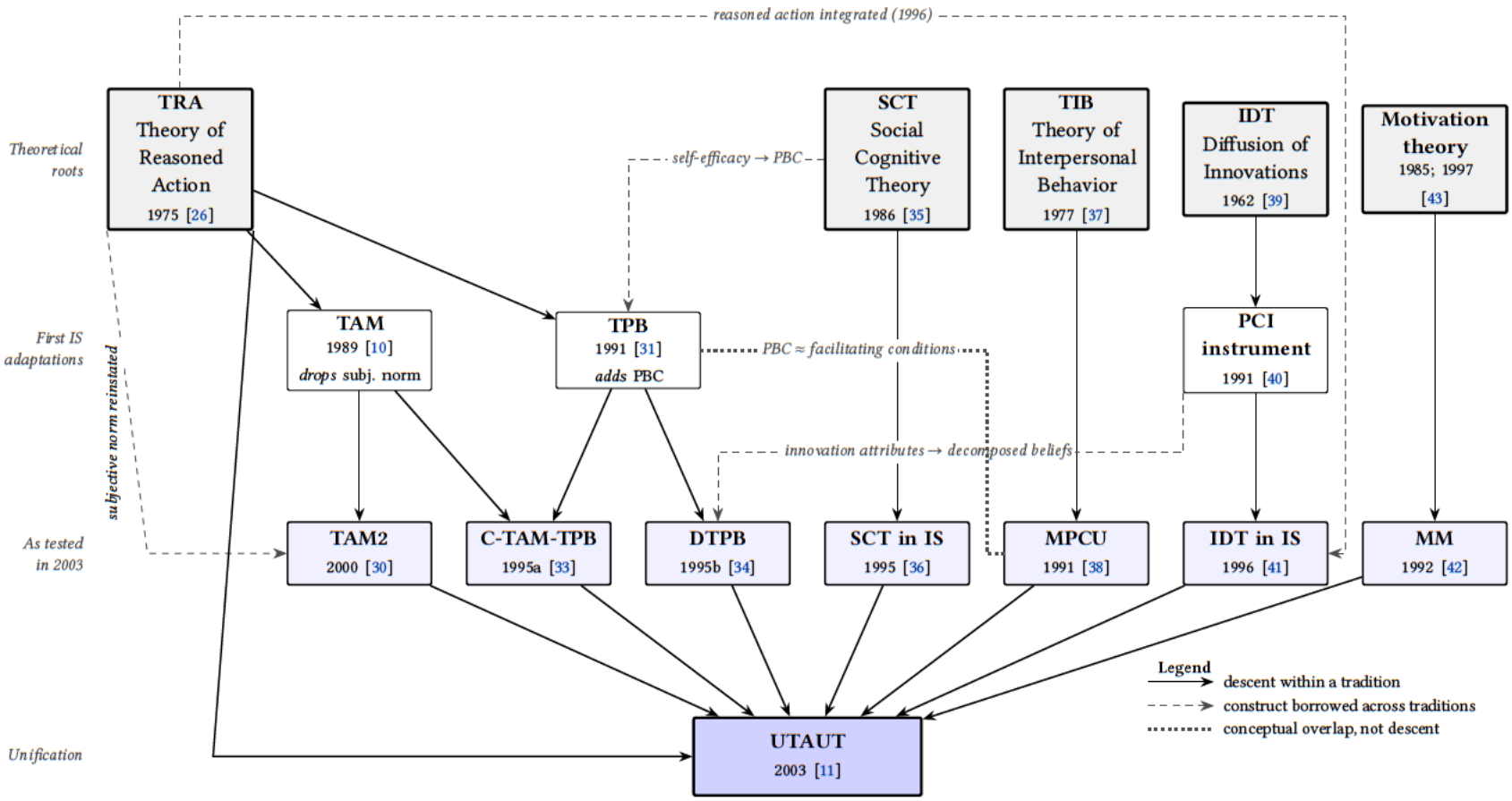

Figure 2.13: Genealogy of the eight models unified by the UTAUT (own illustration based on [11]). Solid arrows denote descent within one research tradition, dashed arrows a construct borrowed across traditions, and the dotted line a conceptual overlap. Where Venkatesh et al. [11] tested a base model together with its extension, the extension is shown (TAM2 for TAM, DTPB for TPB).

Effort Expectancy (EE) "the degree of ease associated with the use of the system" [11]. Based on perceived ease of use, complexity, and ease of use.

Social Influence (SI) "the degree to which an individual perceives that important others believe he or she should use the new system" [11]. Based on subjective norm, social factors, and image.

Facilitating Conditions (FC) "the degree to which an individual believes that an organisational and technical infrastructure exists to support use of the system" [11]. Based on perceived behavioral control, facilitating conditions, and compatibility.

PE, EE, and SI are hypothesized to determine behavioral intention, while FC is hypothesized to affect usage directly, but not intention. The remaining three constructs, computer self-efficacy, computer anxiety, and attitude toward using technology, were hypothesized to have no direct effect on intention:

Computer Self-Efficacy "a judgment of one's capability to use a computer" [36].

Computer Anxiety "an individual's apprehension, or even fear, when she/he is faced with the possibility of using computers" [44].

Attitude Toward using Technology "an individual's overall affective reaction to using a system" [11].

Self-efficacy and anxiety are expected to act indirectly through effort expectancy [44], and the effect of attitude is treated as spurious once PE and EE are included [11]. Additionally, Behavioral Intention (BI) is defined as the person's intention to use the system, which captures "how hard people are willing to try, of how much of an effort they are planning to exert, in order to perform the behavior" [31].

The effects of the core constructs are moderated by gender, age, experience, and voluntariness of use, where voluntariness is defined as "the degree to which use of the innovation is perceived as being voluntary, or of free will" [40]. Age moderates all key relationships [11].

Questionnaire Items For each construct, the four items with the highest loadings from the foundation models were retained and measured on a 7-point Likert scale. The internal consistency was at least 0.77 for all constructs [11]. The original questionnaire, including the non-significant constructs, thus consisted of 31 questions, as BI only used three questions (Appendix A.4). Removing the non-significant constructs results in a questionnaire consisting of just 19 items, the version most commonly used.

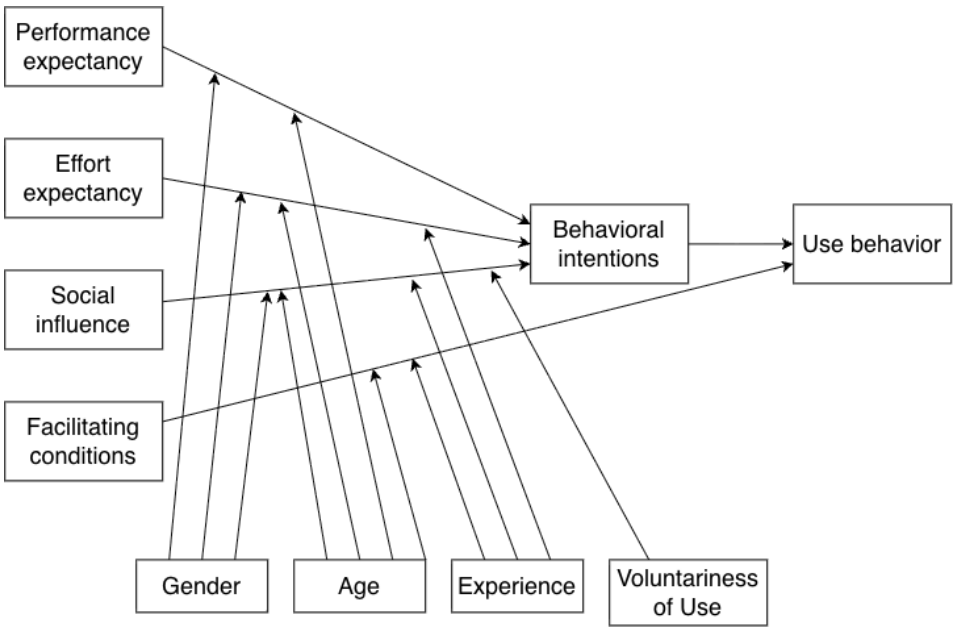

Figure 2.14: Unified Theory of Acceptance and Use of Technology (adapted from [11])

Limitations and Future Work Venkatesh et al. [11] name several limitations themselves, including the preliminary scales, using only standard intention measures, and the model offering little prescriptive guidance for system design. They propose behavioral expectation and habit as directions for future work, while also noting that the high explained variance depends on complex interaction with the moderators [11]. Subsequent reviews have pointed out that these moderators are rarely tested, so the UTAUT holds "only in terms of its main effects" [45].

2.1.3 Post-UTAUT

Since its publication, the UTAUT has been extended in numerous ways, by adding new exogenous predictors (coming from outside a system), new endogenous mechanisms (coming from inside a system), new moderators, or new outcomes [45]. The authors themselves note that this success has created a "UTAUT paradigm" in which incremental extensions dominate while it is "essentially a static theory" [45]. Meta-analyses further show that the original paths are less robust than assumed, with attitude, initially dropped by the UTAUT, partially mediating the effects of all four core constructs [46], and several original paths becoming weak or even negative once additional predictors such as compatibility are introduced [14]. A recurring criticism concerns the treatment of emotions, as acceptance models rely on "naïve and over-simplified notions of affect or emotions" [47], typically merging them into attitude, even though they are different constructs.

2.2 Affective Computing

Affective Computing (AC) is defined as "computing that relates to, arises from, or deliberately influences emotions" [48]. Within HCI, Hudlicka [49] identifies recognizing the user's affect, adapting to it, expressing affect, and modeling affect in user and machine as the four roles of affective systems. Emotions play a part in all goal-directed computer use, as users in a good mood rate both the interface and their work more positively [50], and ignoring the user's emotional state can make an interface appear "cold, socially inept, untrustworthy, and incompetent" [50]. Hudlicka [49] further suggests that affect can serve as "a measure of user satisfaction and thus of interface or product usability".

The terms used in the field of AC are not used consistently, and due to the large number of varying definitions, there is a lot of debate about them [51]. Following Scherer [52], an emotion is "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism". The five components are appraisal, bodily symptoms, action tendencies, motor expression, and subjective feeling [52]. Emotions are short, usually lasting seconds rather than minutes [53], whereas moods are diffuse, longer-lasting states without a specific object that they are directed at [50]. Affect, on the other hand, is used as an umbrella term for these states [49]. Lastly, displayed affect refers to the motor-expression component only, meaning what can be observed in the face or voice. Expression allows emotions to be inferred, but gives no direct access to the subjective feeling, which can only be assessed through self-report measures [52].

2.2.1 Emotion Models

To better assess emotions, different emotion models have been developed that provide a system in which emotions can be quantified. The most prominent models are the categorical emotion model and the dimensional emotion model.

Categorical Emotion Model Categorical models assume a limited number of discrete basic emotions. Ekman [53], the most influential researcher supporting this model, describes basic emotions as discrete states that have "evolved through their adaptive value in dealing with fundamental life-tasks". Ekman [53] names anger, fear, sadness, enjoyment, disgust, and surprise as core basic emotions, each forming a family of related states, and he regards facial expressions as the strongest evidence for distinguishing them. At the same time, Ekman and Friesen [54] note culture-specific display rules governing which expressions are shown.

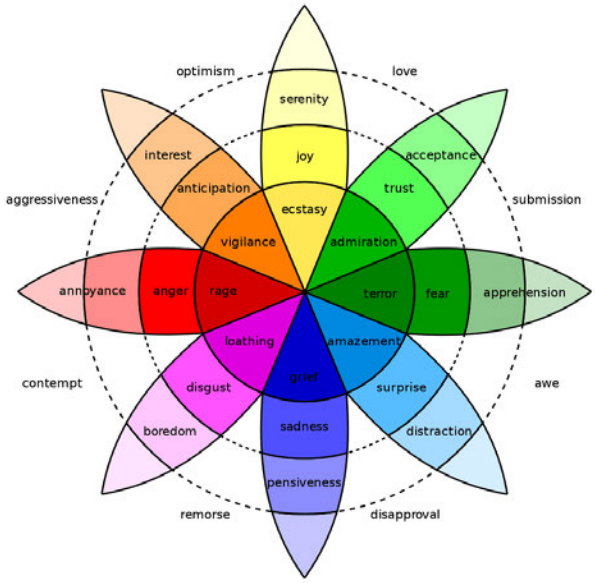

Figure 2.15: Plutchik's wheel of emotion [55]

Plutchik [55], [56] proposed eight primary emotions arranged as four bipolar pairs on a wheel, with similar emotions next to each other, opposites apart, intensity as a third dimension, and complex emotions as mixtures of primary ones (Figure 2.15). Researchers have proposed between three and eleven basic emotions, with there being "no unequivocal way" to settle the number [56].

Dimensional Emotion Model Dimensional models describe affective states as points in a continuous space. An early version by Schlosberg [57] arranged facial expressions on a circular surface defined by pleasantness–unpleasantness and attention–rejection, with level of activation as a third dimension. Mehrabian and Russell [58] proposed three independent, bipolar dimensions (pleasure, arousal, and dominance) as the common core of all emotional responses. Russell [59] showed that affect is better represented by two bipolar dimensions, pleasure–displeasure (valence) and activation–deactivation (arousal), around which emotion concepts are arranged in a circle (Figure 2.16). In a scaling study, eight reference concepts fell in exactly the predicted circular order and close to their hypothesized positions, such as excitement at 45° and depression at 225° [59].

2.2.2 Emotion Recognition Strategies

There are a number of ways to assess either displayed or subjective affect in the field of AC. Each has its own set of advantages and disadvantages. Multimodal fusion aims to leverage the advantages of multiple measures.

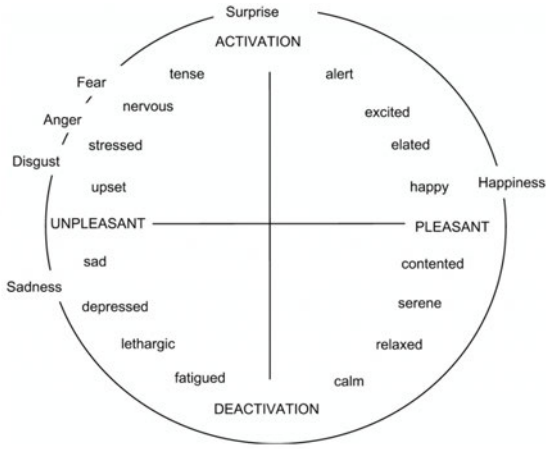

Figure 2.16: Russell's circumplex model of emotions [59]; image: https://psu.pb.unizin.org/psych425/chapter/circumplex-models/

Self-Report Measures Self-report is the only way to access the subjective feeling component of emotion [52] and is considered the most direct measure [50]. The Self-Assessment Manikin (SAM) by Bradley and Lang [60] is a common instrument where users rate pleasure, arousal, and dominance on a 9-point pictorial scale. Another often used instrument is the Positive and Negative Affect Schedule (PANAS), which measures positive and negative affect with ten adjectives each [61]. However, self-report only captures conscious affect, depends on memory, and interrupts the experience it measures [50], while also potentially priming the participants [52].

Behavioral Measures Behavioral measures infer affect from observable expressions of the user. Facial Expression Recognition (FER) via webcam is nonintrusive, reasonably accurate, and cheap [62]. Speech Emotion Recognition (SER) is generally less accurate than FER, with pitch mainly indicating arousal [62]. Text-based approaches classify affect from written language using lexicons or machine learning [62]. A general limitation of all behavioral channels is that people can control them to hide their emotions [63].

Physiological Measures Physiological signals such as Electrodermal Activity (EDA), heart rate, and Electroencephalography (EEG) reflect activity of the central and autonomic nervous system that is "largely involuntarily activated and therefore cannot be easily controlled" [63]. Skin conductance has been used as an index of arousal since Schlosberg [57] and is generally considered better at recognizing arousal than valence [63]. Electrodermal features follow almost the same trend for all emotions [63], and no physiological measure identifies emotion categories reliably [64].

Multimodal Fusion Combining modalities can compensate for the weaknesses of individual affect channels. Fusion can take place at the data, feature, or decision level [62]. A meta-analysis of multimodal affect detection systems found that they consistently outperform their best unimodal counterpart, but only modestly [65].

2.2.3 FER in Practice

Most FER systems build on the Facial Action Coding System (FACS) [66], described as "the most comprehensive and widely used objective taxonomy for coding facial behavior" [67], which decomposes facial expressions into anatomically defined Action Units (AUs). Automated systems can replace manual FACS coding, which requires certified coders and roughly one hour per minute of video [62].

The iMotions research platform uses the AFFDEX SDK by Affectiva [68], [69], [70]. AFFDEX detects faces and facial landmarks, extracts features, and classifies 13 AUs with support vector machines trained on tens of thousands of manually coded images. Emotion scores are derived from AU combinations and range from 0 to 100, and valence from −100 to 100 [67], [70]. In the AFFDEX context, valence is defined as "the overall emotional tone, ranging from negative to positive" and arousal, named engagement in the iMotions software, as "measur[ing] the level of expressiveness and involvement" [69].

Validation studies show that AFFDEX works well for posed, prototypical expressions but less well for natural ones. On standardized picture sets, an earlier version of AFFDEX classified 73 % of expressions correctly, though at times confusing fear with surprise and anger with sadness [71]. Across eight commercial classifiers, however, mean recognition rates were below those of human observers (54 % vs. 72.5 %) and most performed worse on spontaneous than on posed expressions [72]. Barrett et al. [64] concluded that facial configurations are only weakly reliable in indicating emotional states. FER results therefore require careful interpretation and should be considered as displayed affect, not as the participant's actual felt emotional state.

2.2.4 Ethics and Bias

The practice of emotion recognition raises ethical questions beyond measurement validity. Commercial emotion AI is dominated by Basic Emotion Theory, despite its scientific basis being contested, and Stark and Hoey [51] argue that predicting an individual's emotional state is "always suspect" and that systems handling emotion data deserve "heightened scrutiny". Facial analysis is also demographically biased, with higher gender misclassifications of

darker-skinned women [73] and FER validation studies have mostly used white faces [71]. Additionally, cultural display norms influence how openly emotions are expressed [54].

2.3 User Experience & Usability

This section introduces the concepts of usability (Section 2.3.1) and user experience (Section 2.3.2), discusses how they relate to each other (Section 2.3.3), and presents standardized questionnaires for assessing perceived usability and UX (Section 2.3.4).

2.3.1 Usability

The term *usability* was coined in the early 1980s to replace the term "user friendly", which had acquired "a host of undesirably vague and subjective connotations" [74]. In practice, it is still often equated with ease of use [75].

According to ISO 9241-11:2018, usability is defined as:

> "[...] the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [76]

However, usability is not an inherent property of a system, product, or service itself but rather relates to the outcome of interacting with said system, product, or service [76]. It is relevant to regular use, learning to use, minimizing risk, and maintenance of systems, products, or services among many other points of contact [76].

ISO 9241-11 was first published in 1998 [77] and revised in 2018 [76]. The revision extended the scope from just products to systems, products, and services, treated usability as an outcome of use rather than a product attribute, and defined satisfaction to include users' emotional responses [76]. Both versions highlight that usability depends on the context of use, influenced by the users, their tasks, equipment, and environment, which is why it must be specified and measured within that context [76], [77]. Bevan [75] also distinguishes usability from a designable product property and finds that it can only be validated through user-based evaluation in a realistic context.

Perceived Usability Usability can be measured objectively, through effectiveness and efficiency metrics such as success rate (effectiveness) and time on task (efficiency), and subjectively through users' satisfaction ratings. In an early review of 180 studies, Hornbæk [78] found that the measurement of satisfaction needed standardized questionnaires and a clear separation of subjective and objective measures. These types of measures, though related, are not interchangeable, as Sauro and Lewis [79] found in 90 usability tests that the correlations among these metrics were strong at the task level (r = .44 to .60) but weak when satisfaction was measured once after the test (r = .16 to .24), and their factor analysis showed an objective and a subjective factor underlying a general usability construct. Similarly, the System Usability Scale (SUS) correlates with other subjective measures at 0.50 to 0.96 but with task success at only 0.22 to 0.50 [19]. Lewis [80] therefore defines *perceived usability* as the subjective component that, along with efficiency and effectiveness, makes up the classical ISO definition of usability, "which is in turn a fundamental component of user experience". This perceived component is closely related to technology acceptance, where Lah et al. [18] found a strong correspondence between perceived ease of use of a version of the TAM and the SUS in regression models.

2.3.2 User Experience

User Experience (UX) emerged as a counter-movement to the more task-focused usability paradigm [81]. Hassenzahl and Tractinsky [81] identify going beyond the instrumental, utilizing emotion and affect, and using the experiential as three threads in UX research. They define UX as "a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, etc.)" [81]. A large-scale survey among 275 UX researchers and practitioners found agreement that UX is dynamic, context-dependent, and subjective, but no consensus on its scope [82]. The definition by Hassenzahl and Tractinsky [81] was the most preferred one.

Another promising comprehensive definition of UX is provided by the ISO 9241-210 standard from 2019:

> "[...] perceptions and responses of users that result from the use and/or anticipated use of a system, product, or service" [83]

A central distinction in UX research is between pragmatic and hedonic quality [84]. Pragmatic attributes (utility and usability) support users in achieving behavioral goals, whereas hedonic attributes emphasize individuals' psychological well-being through stimulation, identification, and evocation [84]. Both are perceived independently and contribute independently to the

overall evaluation of a product [84], [85]. Diefenbach et al. [85] showed that hedonic quality has become an established UX concept, often defined in contrast to the pragmatic.

The Components of User Experience (CUE) model by Thüring and Mahlke [86] integrates these aspects (Figure 2.17). In the CUE model, interaction characteristics influence the perception of instrumental qualities ("the experienced support the system provides and the ease of its use") and non-instrumental qualities ("the look and feel of the system") [86]. Together with emotional reactions, these characteristics determine the overall appraisal of a system [86]. Empirically, a system with poor usability led to lower self-reported valence, higher arousal, and higher electrodermal activity [86], linking usability to the affective measures in this thesis.

2.3.3 Relationship between UX and Usability

Although the relationship between UX and usability is an important factor in HCI, finding a widely accepted differentiation between the terms has not been an easy task, as Hassan and Galal-Edeen [87] set out to find out if UX is a part of usability or if usability is a part of UX. They conclude that neither user experience nor usability alone can determine if the product meets users' needs and goals and state that:

> "[U]sability can be seen as a subset of user experience and we consider it as the heart of user experience." [87]

Bevan [88] argues that how UX is defined determines what is measured, as UX can be seen as an extension of the satisfaction component of usability, as something separate from performance-oriented usability, or as an overarching term for all user perceptions and responses. Regardless of terminology, the objective remains to optimize human performance and user satisfaction by achieving both pragmatic and hedonic goals [88]. Law et al. [82] conducted a survey in which the participants agreed that usability is a necessary precondition for good UX. In this thesis, perceived usability is therefore understood as the instrumental or pragmatic subset of UX, in line with Lewis [80] and Hassan and Galal-Edeen [87].

2.3.4 Assessing Perceived Usability and UX

The SUS, published by Brooke [89] in 1996 as a "quick and dirty" method to assess usability, is among the most widely used standardized questionnaires for measuring perceived usability, with it accounting for more than 40 % of post-study questionnaire usage in industrial usability studies [79]. It consists of 10 items with an alternating tone, meaning alternate positive and negative wording, rated on a 5-point Likert scale (Appendix A.2). Initially developed to fill the gap in subjective usability in the 1980s and 1990s, it has been used extensively due to it being

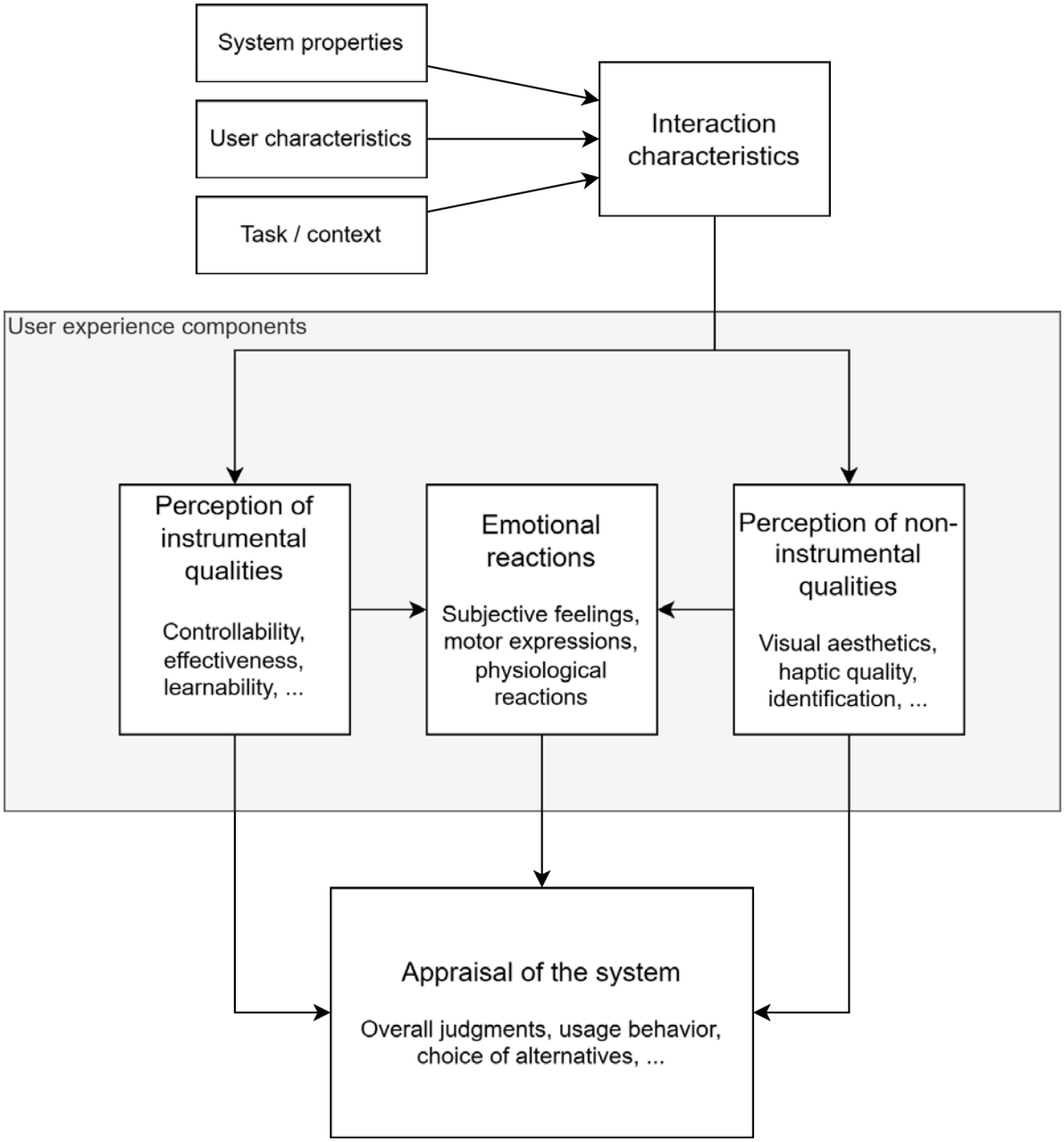

Figure 2.17: The CUE model (components of user experience) [86]

free to use, providing a single easy-to-interpret score from 0 to 100, and its quick-to-fill-out nature [19]. The SUS has been found to be a highly robust and versatile tool, having been used in hundreds of studies [90] and even surviving the change from an alternating to a positive tone [91]. In an attempt to reduce the number of user mistakes when filling out the questionnaire, Sauro and Lewis [91] created a positive-tone version of the SUS. Although they did not find a significant difference between the mixed-tone SUS and the positive-tone SUS, they recommend using the positive-tone version, as it reduces the risk of mistakes from both the users while filling out and the researchers while transcribing.

In order to collect data from non-English-speaking German users, a German translation of the SUS is required. Brix et al. [92] set out to compare different available German translations of the SUS and found the mixed-tone version of Rummel [93] to be "[...] a good compromise between wording and methodically clean development". Perrig et al. [94] support this in a large study, comparing different variations of German translations and recommending the use of either their presented positive-tone variation of the SUS translation based on the work by Rummel [93], or the original Rummel [93] translation.

In order to interpret the SUS score, a proper understanding of its meaning is necessary. Bangor et al. [90] conducted an empirical evaluation of the SUS and concluded that a score of 70 and above is considered acceptable, with scores of 80 being strongly above average [95].

Sauro and Lewis [95] offer a new look at the interpretation of SUS scores by creating a grading scale that allows for intuitive individual assessment of scores and comparison between scores. Table 2.2 shows the full grading scale.

Table 2.2: The Sauro-Lewis curved grading scale (CGS) [95]

SUS score range	Grade	Percentile range
84.1–100	A+	96–100
80.8–84.0	A	90–95
78.9–80.7	A–	85–89
77.2–78.8	B+	80–84
74.1–77.1	B	70–79
72.6–74.0	B–	65–69
71.1–72.5	C+	60–64
65.0–71.0	C	41–59
62.7–64.9	C–	35–40
51.7–62.6	D	15–34
0.0–51.6	F	0–14

Finstad [96] developed the Usability Metric for User Experience (UMUX) as a subjective assessment of an application's usability with the goal of producing similar results to the SUS.

It is organized around the ISO 9241-11:1998 standard [77] and consists of four items that are scored on a 7-point Likert scale. Finstad [96] found a high correlation of $r = 0.96$ between the UMUX and the SUS, indicating that they assess the same construct. Distilling the UMUX down to two items, Lewis et al. [97] created the UMUX-LITE, an even shorter alternative that correlated with the SUS at $r = 0.81$.

In 2013, Kortum and Bangor [98] published an extensive collection of quantitative usability ratings of 14 everyday products and product categories to help judge how new products compare to well-known products. They used the SUS in an online study with more than 1000 participants, capturing their experiences interacting with a variety of products. Building on their work and attempting to compare different usability measures, Lewis [80] replicated the study with a selection of four everyday products in 2019. His results, displayed in Table 2.3 with both the raw scores and the grades based on the Sauro-Lewis grading scale [95], are consistent with previous research, and he concluded that the alternative non-SUS measures seem to be acceptable for use in practice. Table 2.4 provides an overview of characteristics of often used usability measures.

Table 2.3: Perceived usability ratings of everyday products [80], with grades from the curved grading scale (Table 2.2)

Measure	Excel	Word	Amazon	Gmail
SUS	69.6 (C)	75.5 (B)	84.8 (A+)	78.0 (B+)
UMUX	71.9 (C+)	76.4 (B)	86.4 (A+)	77.7 (B+)
UMUX-LITE	74.0 (B−)	78.0 (B+)	86.6 (A+)	77.7 (B+)
CSUQ	68.7 (C)	72.8 (B−)	82.3 (A)	75.2 (B)

The UMUX-LITEr is a regression-adjusted version of the UMUX-LITE whose scores are designed to match the SUS [99]. Because the regression adjustment restricts the UMUX-LITEr to a range of 22.9 to 87.9, it underestimates products with high usability. For products rated B+ or better, the unadjusted UMUX-LITE is therefore recommended [18], [80].

Lewis [100] analyzed how the number of response options affects UMUX-LITE scores and concluded that the number of response options does not matter much and recommends using a 7-point Likert scale.

Lewis [19] notes a close correspondence of the UMUX, the UMUX-LITE, the UMUX-LITEr, and the Computer System Usability Questionnaire (CSUQ) with the SUS.

Still, Lewis [19] strongly recommends the use of the SUS for usability research, as it is likely to remain in the position of being the most widely used measure of perceived usability. He argues that for unsupervised or remote settings, the positive-tone version should be used.

Table 2.4: Correlation of different usability measures with the SUS

Name of Measure	Number of Items	Tone	Scale	Correlation with SUS
UMUX [96]	4	Mixed	7-point Likert	0.96
UMUX-LITE [97]	2	Positive	7-point Likert	0.81 / 0.85
UMUX-LITEr [99]	2	Positive	7-point Likert	0.83
CSUQ [80], [101]	16	Positive	7-point Likert	0.76

Lewis [100] highlights the need for more usability studies, as much of the SUS and UMUX-LITE literature was collected using general use surveys rather than direct post-task questionnaires. Lah et al. [18] similarly found that the variance explained by perceived usability depended on the context of use, with mandatory use yielding the highest and voluntary use the lowest, calling for replications with other systems to improve generalizability.

In practice, perceived usability is often reported together with the Net Promoter Score (NPS), which is based on the single question "How likely is it that you would recommend [x] to a friend or colleague?" on an 11-point scale. Respondents scoring 9–10 are considered promoters, and those scoring 0–6 are considered detractors. The NPS is the percentage of promoters minus the percentage of detractors [102]. The SUS correlates with this item and explains 39 % of its variance [19].

For the broader UX context, the User Experience Questionnaire (UEQ) is a widely used instrument [103]. It consists of 26 items on 7-point semantic differentials grouped into six scales. These scales include attractiveness, three pragmatic scales (perspicuity, efficiency, dependability) and two hedonic scales (stimulation, novelty) [103]. Its pragmatic scales correlated with task completion time, while its hedonic scales did not [103].

Internal consistency is commonly assessed with Cronbach's coefficient alpha [104], with values between 0.70 and 0.95 being considered acceptable for research [19].

2.4 Large Language Models

This section outlines the Transformer architecture on which current LLMs are based (Section 2.4.1), how pre-trained models are aligned with user intent (Section 2.4.2), their main limitations and how RAG addresses them (Section 2.4.3), and the model families relevant to the assistants developed in this thesis (Section 2.4.4).

2.4.1 Architecture

LLMs build on the Transformer architecture [105], which replaces recurrence with self-attention. This is a mechanism "relating different positions of a single sequence in order to compute a representation of the sequence" [105]. As every position can attend to every other in a single step, Transformers capture long-range dependencies and can be trained in parallel [105]. Most current LLMs are decoder-only models pretrained in a self-supervised way to predict the next token on large text bodies [106].

The performance of such models improves with their scale so that language-modeling loss follows smooth power laws in the number of parameters, the amount of training data, and the training compute, with only weak dependence on architectural details [21]. However, model size is not the only factor, with Hoffmann et al. [107] showing that model size and training data should be scaled in equal proportion by having their 70B-parameter model, trained on 1.4 trillion tokens, outperform the larger Gopher model (280B) at the same compute budget.

2.4.2 Alignment

Pre-trained language models predict text but do not necessarily follow user instructions. Ouyang et al. [108] found that scaling the model parameters alone does not automatically solve this, as the fine-tuned 1.3B-parameter model in their study was preferred over the 175B-parameter GPT-3. Alignment typically combines supervised fine-tuning on instruction–response pairs with reinforcement learning from human feedback, with reward models being trained on human preference rankings [108]. In the case of Ouyang et al. [108], the goal is to achieve a model that is helpful, honest, and harmless. Direct Preference Optimization achieves the same objective without a separate reward model or reinforcement learning loop [109] and is used by model families such as Llama 3 [110].

Aligned LLMs can be adapted to tasks through prompting rather than retraining, as shown in in-context learning where the task is specified purely through text, either as an instruction (zero-shot) or with a number of demonstrations (few-shot), without changing any of the model weights [111]. Chain-of-thought prompting, providing "a coherent series of intermediate reasoning steps that lead to the final answer", further improves performance on multi-step problems, but only in sufficiently large models starting at around 100B parameters [112].

2.4.3 Limitations and Retrieval-Augmented Generation

Despite their linguistic fluency, LLMs have well-known limitations, including hallucinations, the "generation of content that is nonsensical or unfaithful to the provided source" [113],

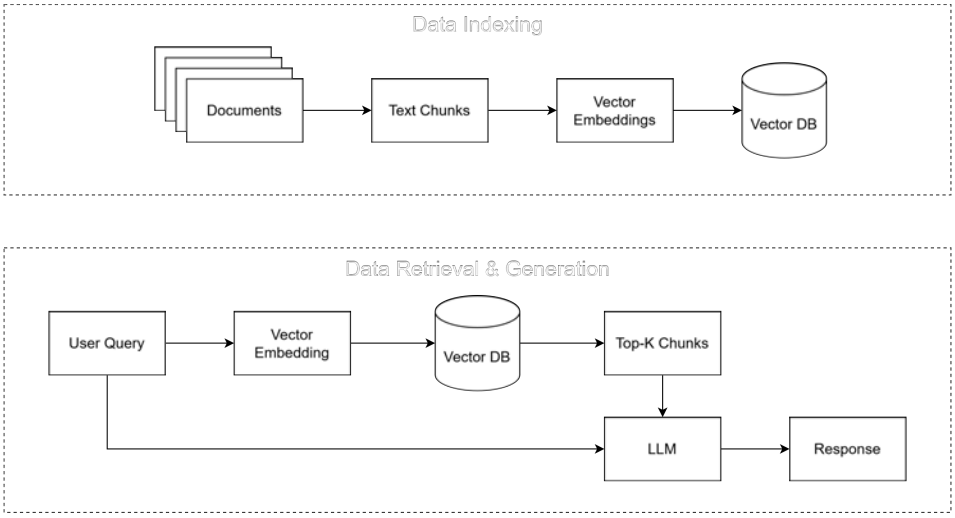

Figure 2.18: RAG pipeline with indexing, retrieval, and generation

which are important to consider in chatbot usage. Ji et al. [114] distinguish intrinsic hallucinations, which contradict the source, from extrinsic ones, which can be neither supported nor contradicted by it. Causes include biases in the training data, decoding strategies that trade faithfulness for diversity, and the tendency of models to prioritize their parametric knowledge over the provided input [114]. Instruction tuning can reduce but not completely remove the problem, as shown by Ouyang et al. [108], who compared the hallucination rate of InstructGPT (21 %) with GPT-3 (41 %). Further limitations include the knowledge cut-off of the training data and the lack of access to external or up-to-date information [113].

These limitations are especially problematic because of the human tendency to perceive fluent model output as coherent and intentional, potentially leading to automation bias [115]. RAG addresses hallucination and outdated knowledge by retrieving relevant data from an external knowledge base and providing it to the model as additional context [116], [117]. Shuster et al. [118] found that in knowledge-grounded dialogue, retrieval substantially reduced hallucination without impairing conversational ability.

Figure 2.18 shows an exemplary RAG pipeline consisting of the data indexing, as well as the retrieval and generation processes.

2.4.4 Model Overview

The number of parameters strongly affects the cost of running an LLM, with inference energy depending on model size, output length, and hardware. A medium-sized LLM uses more than 40 times the electricity of a small 1.7B-parameter model for the same text task [23], and multi-purpose generative models consume orders of magnitude more energy per query

than task-specific ones [22]. Across models, per-prompt energy varies by more than two orders of magnitude, with DeepSeek-R1 consuming 19.251 Wh of electricity, and Llama 3.1-8B only consuming 0.052 Wh with the same prompt settings [119]. This has led to growing interest in Small Language Models (SLMs), as they offer advantages in efficiency, accessibility, customization, and privacy, since they can be run locally without relying on model providers and sending sensitive data over networks [120]. Some authors argue that SLMs are "sufficiently powerful" for most uses in agentic systems [121], though they still lag behind on complex reasoning [122] and can struggle with complex multi-step queries [123].

Model families are often released in several sizes with openly available weights, such as the Llama 3 family with 8B, 70B, and 405B parameters [110], or the Swiss Apertus model family [124]. Apertus was initially released as an 8B and 70B model family, but since its release, 0.5B, 1.5B, and 4B models have followed [125]. These smaller variants are particularly suitable for local hosting.

3 Related Work

This chapter presents current research on the roles of affective computing and usability in the technology acceptance of AIVAs (Section 3.1), and gives some insight into recent developments of AIVAs (Section 3.2). Additionally, research gaps tackled by this thesis are derived (Section 3.3).

3.1 Acceptance of AIVAs

This section reviews how affect (Section 3.1.1) and usability (Section 3.1.2) have been integrated into technology acceptance research, introduces the UVA-UTAUT, which combines both (Section 3.1.3), and summarizes findings on the acceptance of AI-based assistants (Section 3.1.4).

3.1.1 Affect and Technology Acceptance

Although early acceptance models mostly use cognitive constructs, the role of affect has been partially explored. In the MPCU, self-reported affect toward PC use had no significant effect on utilization [38], and in the UTAUT, the affective construct, attitude toward using technology, was dropped because its effect disappeared once PE and EE were controlled for [11]. Other studies, however, found affective constructs to matter, as perceived enjoyment predicted the intention to use computers beyond perceived usefulness [42], and computer anxiety and enjoyment shaped perceived ease of use [44], which led to their inclusion in TAM3 [29].

Only a few studies used a more comprehensive emotion model like the Consumer Acceptance of Technology (CAT) model, which extended the TAM with pleasure, arousal, and dominance dimensions [126]. Kulviwat et al. [126] found that pleasure and arousal significantly predicted attitude and that the model explained 53 % of the variance in intention compared to 38 % for the TAM. Beaudry and Pinsonneault [16] showed that discrete emotions such as anxiety, excitement, and happiness affect IT use, mostly indirectly through coping behaviors. For AI-based devices, Gursoy et al. [12] proposed the AIDUA model, in which performance and

effort expectancy evoke an emotional response that predicts the willingness to use, and Sohn and Kwon [13] found enjoyment to be the strongest predictor of intention to use AI-based intelligent products.

Although these studies take a deeper look into the role of affect, all of them rely on self-report questionnaires. None of them use implicit behavioral measures of the displayed affect of users, even though self-report only captures conscious affect and depends on memory [50].

3.1.2 Usability and Technology Acceptance

Usability and technology acceptance are conceptually close, as the UMUX-LITE, a measure of perceived usability, was constructed from items mapping onto the TAM's perceived usefulness and ease of use constructs [97]. Responding to doubts about the value of further usability research, Lah et al. [18] mapped out the relationship between perceived usability and the PU and PEOU constructs of a modified version of the TAM by Davis [10]. They found a significant relationship between perceived usability and PEOU, suggesting that PEOU of the modified TAM appears to be a measure of the construct of perceived usability [18]. Mlekus et al. [127] took a similar approach by using UX qualities as additional constructs of the TAM. Perspicuity strongly predicted perceived ease of use, and output quality predicted perceived usefulness [127]. Studies like these are not common, and Lewis [19] argues that future research on the SUS should also examine its relationship with technology acceptance measures.

It is debated whether general usability questionnaires are suitable for conversational agents, with Borsci et al. [128] arguing that general satisfaction scales miss the conversational aspects of chatbot interaction. Holmes et al. [129] developed the Chatbot Usability Questionnaire (CUQ) and found high correlations between it, the SUS, and the UEQ. In their data, the CUQ was more strongly linked to the UEQ than to the SUS, suggesting that the CUQ captures a broader UX construct, whereas the SUS focuses on perceived usability. Borsci et al. [128] also developed the Chatbot Usability Scale (BUS-15), which was later shortened to the BUS-11 [130], and validated it in several languages. The CUQ has not been fully validated, although promising exploratory factor analyses have been conducted [131], and the BUS has so far only been validated for customer-service chatbots with an unstable factor structure [132]. Furthermore, Følstad et al. [133] note a general "lack of common definitions, metrics and validated scales" in chatbot research.

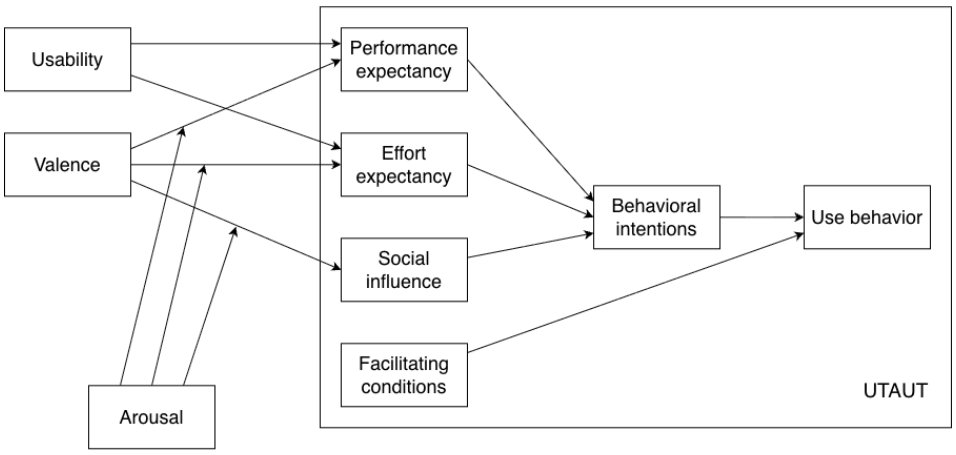

Figure 3.1: Usability-Valence-Arousal UTAUT (adapted from [20])

3.1.3 The UVA-UTAUT

Hildebrandt et al. [20] proposed the Usability-Valence-Arousal UTAUT (UVA-UTAUT, Figure 3.1) for AIVAs in the context of housing modernization. The model extends the UTAUT with perceived usability as a predictor of PE and EE, and with displayed valence, moderated by arousal, as a predictor of PE, EE, and SI.

The choice of the valence–arousal model [59] was motivated by the limited research on affect in technology acceptance settings, which mostly used isolated aspects of emotions [20]. The paper presented a study protocol combining the UTAUT questionnaire, the SUS, and implicit FER measures.

Hildebrandt et al. [25] evaluated the model in a pilot study with six participants, in which participants planned a modernization measure on a web platform with an integrated AIVA while their facial expressions were analyzed with AFFDEX. Due to the limited sample size, they reported effect sizes and minimum sample size estimates instead of significance tests. Perceived usability showed large effects on PE and EE, which were estimated to be detectable with 11 and 9 participants, respectively. Of the affective pathways, only the valence × arousal interaction on PE reached a large effect (estimated minimum of 17 participants), whereas the interaction was small for EE and negligible for SI.

With six participants, the pilot effects are only trends, rather than reliable estimates, and it is unclear whether they stay consistent in a fully powered study. Additionally, the model has only been evaluated for a single application, meaning that generalization is unknown. Finally, it has not been tested whether perceived usability relates more strongly to EE than to PE, as the conceptual constructs would suggest [18].

3.1.4 Acceptance of AI-Based Assistants

For personal intelligent agents such as Siri or Alexa, perceived intelligence was a strong antecedent of perceived usefulness, ease of use, and trust, while perceived anthropomorphism increased enjoyment, with usefulness and enjoyment predicting intention to use [134]. For digital voice assistants, perceived usefulness, trust, and rapport directly predicted acceptance [135]. For LLM-based chatbots, early studies on ChatGPT found PE to dominate, whereas EE was uniformly positive, as such tools require little effort to use [136]. A systematic review of 60 studies on AI acceptance confirmed that PE and EE positively predict intention across domains and recommended adding trust and attitudes to the traditional models [137].

3.2 Research on AIVAs

Conversational agents have a long history in HCI, with ELIZA [138] being one of the first. It scanned user input for keywords and transformed the sentence according to attached rules, without any representation of meaning [138]. Some users were hard to convince that the program was not human, which led Weizenbaum [138] to conclude that the credibility attributed to machine output demands investigation. Later systems such as PARRY and ALICE refined this pattern-matching approach, before voice assistants like Siri and Alexa brought conversational interfaces into everyday use [139]. Pattern matching produces fast but repetitive answers, whereas machine-learning approaches can use the whole dialogue context but require large amounts of training data [140]. LLMs (Section 2.4) continue this development, as they generate free-form answers instead of selecting predefined ones, at the price of the hallucinations described in Section 2.4.

For HCI, this shift means that the conversation itself becomes the object of design [141]. In an early survey of chatbot users, productivity was the main motivation for use (68 %), ahead of entertainment (20 %), social motives (12 %), and novelty (10 %) [142]. A systematic review of 83 user studies on text-based chatbots by Rapp et al. [17] identified acceptance, UX, conversational issues, emotions, and humanness as the main research themes. The review found negative emotions to be the main reason for abandoning a conversation and criticized the field as fragmented, with poorly defined constructs and only four studies reporting a power analysis [17]. Følstad et al. [133] also describe emotion detection through non-verbal behavior as "of high relevance to chatbot user experience and design". Beyond conventional usability concerns, AI-based systems also raise specific adoption barriers, such as opaque models, model errors that erode trust over time, and algorithm aversion [15]. The relevance of these factors may change after adoption, as enjoyment no longer predicted the continued use of personal intelligent agents among experienced users [143].

LLM-based assistants have only recently been applied to energy-related building decisions. Ghani et al. [144] presented a GPT-4o-based RAG assistant for Swedish housing cooperatives, whose answers closely matched those of energy advisors (47 of 50 answers reached a cosine similarity of at least 0.8) and which answered 85 % of building-specific questions correctly. Shu and Zhao [145] benchmarked seven LLMs on retrofit decisions for 400 homes. The models performed considerably better when minimizing CO_2 emissions than when minimizing the payback period (top-1 accuracy of 38.8 % vs. 11.1 %) and agreed with each other less than expected by chance. Both studies evaluate the generated answers technically, without involving users. Whether people accept such assistants, and which factors shape this acceptance, remains unanswered.

3.3 Research Gaps

Based on the provided overview of current related work, the following research gaps are identified:

1. **Affect is under-represented and only self-reported.** Acceptance models either ignore affect or reduce it to single constructs such as anxiety or enjoyment [47], and in cases where emotions are considered, they are measured through retrospective questionnaires. The relationship between the affect users display during the interaction and their acceptance has, to the best of the author's knowledge, not been examined beyond the pilot study of the UVA-UTAUT [25].

2. **The link between perceived usability and acceptance is barely studied.** Only a few studies related standardized usability measures to acceptance constructs [18], [127], and this relationship has been explicitly stated as an open question [19].

3. **AI assistants require context-specific acceptance research.** Classic acceptance models may not be applicable to AI technologies [12] and the UTAUT is "less robust than it is often assumed" and should be tested for chatbots [14].

4. **LLM-based assistants in the building domain lack user evaluation.** Existing work evaluates LLM answers for energy advice technically against expert responses [144], [145], but not how users perceive and accept such assistants.

5. **The conversation design of an AIVA is not compared.** With conversational interfaces, the conversation itself becomes the object of design [141], yet acceptance studies evaluate one given assistant at a time [136], [137], and the UVA-UTAUT has likewise only been examined for a single application [25]. Whether an assistant that guides the user through a task is perceived and accepted differently from an assistant

that answers questions on demand alongside a conventional interface has, to the best of the author's knowledge, not been examined for the same task.

6. **The effect of model size on users is unknown.** Models of different sizes are compared on benchmarks and through ratings of their outputs [21], [108], [145], and SLMs are discussed in terms of efficiency, privacy, and reasoning capability [120], [122]. Whether the size of the underlying LLM changes how users perceive and accept an assistant has not been studied, although smaller models require only a fraction of the energy [119] and can be hosted locally [120].

4 Research Design

This chapter derives the hypotheses for the research questions stated in Chapter 1 to further explore the previously established research gaps (Section 4.1) and presents the concise research plan followed in this thesis (Section 4.2).

4.1 Hypotheses

Each research question is operationalized by one group of hypotheses: H1 addresses perceived usability (RQ1), H2 displayed affect (RQ2), H3 the application (RQ3), and H4 the model size (RQ4).

Perceived Usability and Displayed Affect (RQ1, RQ2) Building on the UVA-UTAUT, which was proposed by Hildebrandt et al. [20] and tested in a pilot study by Hildebrandt et al. [25], the initially suggested hypotheses, labeled following the scheme of Mlekus et al. [127], are slightly altered as follows:

H1a: The *perceived* usability of an AIVA is positively related to its performance expectancy.

H1b: The *perceived* usability of an AIVA is positively related to its effort expectancy.

H1c: The relationship between perceived usability and effort expectancy is stronger than that between perceived usability and performance expectancy.

H2a: Displayed valence during AIVA use is positively related to performance expectancy, and this relationship is moderated by arousal, such that the effect is stronger at higher levels of arousal.

H2b: Displayed valence during AIVA use is positively related to effort expectancy, and this relationship is moderated by arousal, such that the effect is stronger at higher levels of arousal.

H2c: Displayed valence during AIVA use is positively related to social influence, and this relationship is moderated by arousal, such that the effect is stronger at higher levels of arousal.

These hypotheses form the foundation of the research conducted within this thesis. Since the SUS captures users' subjective assessment rather than objective effectiveness and efficiency [80], the construct measured is *perceived usability* rather than *usability*, as previously used in the work of Hildebrandt et al. [20] and Hildebrandt et al. [25].

Effort expectancy is rooted in perceived ease of use and complexity [11], with perceived ease of use in a post-use TAM measure found to measure what is estimated to be the same construct as perceived usability [18]. Therefore, a positive relationship between perceived usability and effort expectancy is expected (H1b). Performance expectancy, on the other hand, is rooted in perceived usefulness [11], and perceived ease of use has been shown to affect perceived usefulness, growing with experience [29]. Additionally, the UMUX-LITE, consisting of only two items, assesses perceived usability through one usefulness-related and one ease-of-use-related item [97], and in the UX-TAM, perspicuity and dependability predicted ease of use, which in turn predicted perceived usefulness [127]. Thus, perceived usability is also expected to relate positively to performance expectancy (H1a), although to a lesser extent than effort expectancy (H1c), as the SUS correlated more strongly with perceived ease of use than with perceived usefulness [18].

Brave and Nass [50] found that users in a positive mood rated both the interface and their work more favorably, and Kulviwat et al. [126] showed that in the CAT model, pleasure and arousal predicted attitude toward a technology. In the AIDUA model, emotions are closely connected to performance and effort expectancy, although there the expectancies precede the emotional response [12], and in the TAM3, enjoyment and anxiety, among others, are determinants of perceived ease of use [29]. Displayed valence is therefore expected to relate positively to both performance expectancy (H2a) and effort expectancy (H2b). Social influence is included, as positive affect may also shape the peer perception of interacting with technology (H2c). Arousal can be viewed as the intensity dimension of emotion [57]; hence, arousal is expected to moderate the valence construct in such a way that the effect of valence will be stronger at higher levels of arousal.

Application and Model Size (RQ3, RQ4) The two developed applications follow different approaches to solving the same technical task: Application A, the Guided Assistant, leads the user through the task in a text-based natural language dialogue, whereas Application B, the Basic Chat, supports a more traditional web form with an assistant chatbot that answers questions (Chapter 5). Differences in perceived usability, displayed affect, and technology acceptance between the applications are therefore expected. The direction of these differences is not specified in advance. A guided dialogue may reduce the effort of a complex procedure, but it also restricts the user's control and depends on the LLM correctly interpreting every input, whereas the form is a familiar interaction pattern in which the chat is only used on demand.

Ouyang et al. [108] found that outputs of a fine-tuned 1.3B model were preferred over those of the 175B GPT-3 model, showing that model size alone does not determine how well a model follows users' intent. Larger models generally perform better on benchmarks [21], but since both Apertus models used in this thesis were post-trained with the same procedure [146], it remains open whether the additional parameters still make a noticeable difference to users once alignment is held constant. For this reason, the hypotheses on model size are likewise formulated without a direction.

In these hypotheses, *displayed affect* refers to the displayed valence and arousal, and *technology acceptance* to the five measured UTAUT constructs: performance expectancy, effort expectancy, social influence, facilitating conditions, and behavioral intention.

H3a: Perceived usability differs between the Guided Assistant (Application A) and the Basic Chat (Application B).

H3b: Displayed valence and arousal during use differ between the Guided Assistant (Application A) and the Basic Chat (Application B).

H3c: The UTAUT constructs differ between the Guided Assistant (Application A) and the Basic Chat (Application B).

H4a: Perceived usability differs between an AIVA based on the small (8B) and on the medium-sized (70B) LLM.

H4b: Displayed valence and arousal during use differ between an AIVA based on the small (8B) and on the medium-sized (70B) LLM.

H4c: The UTAUT constructs differ between an AIVA based on the small (8B) and on the medium-sized (70B) LLM.

H3b, H3c, H4b, and H4c each cover several measures. They are evaluated per measure, and a hypothesis is considered partially supported if the difference is found for only some of its measures.

4.2 Research Plan

To evaluate the hypotheses and research questions, two different AIVAs were developed as part of the work conducted in the IntelMOD project [8]. These applications are designed to assist the user in conducting a building modernization measure, using either a small (8B) or medium-sized (70B) LLM. The whole design and development cycle is presented in Chapter 5. In a laboratory user study with a 2×2 factorial design (application within subjects, model size assigned per task), participants use both applications while their facial expressions are

recorded and are then prompted to fill out questionnaires on their perceived usability and technology acceptance (Figure 4.1; Chapter 6). H1a–H2c are tested with regression analyses per application, and H3a–H4c with linear mixed-effects models of the two design factors (Chapter 7).

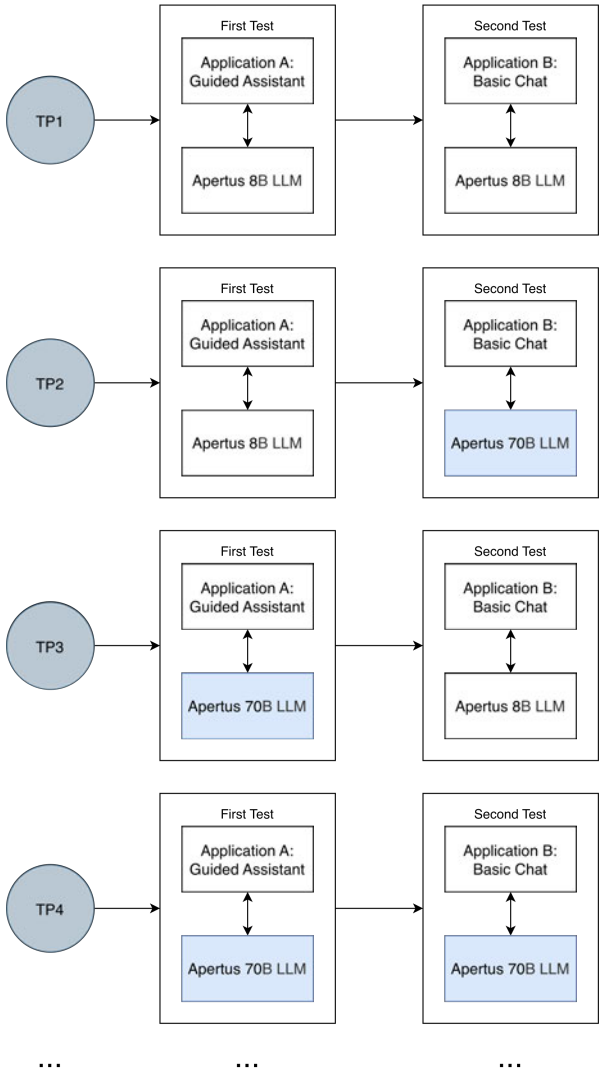

Figure 4.1: Task sequences of the first four participants (TP1–TP4)

5 System Design and Implementation

This chapter describes the research project context and the resulting design goals for the AIVAs (Section 5.1), as well as the functional and non-functional requirements (Section 5.2) and further design constraints (Section 5.3). In addition to the system architecture (Section 5.4), an overview of the RAG configuration (Section 5.5) and the technical validation (Section 5.6) is provided.

5.1 Context and Design Goals

The AIVAs were developed in the joint research project *IntelMOD*, funded by the German Federal Ministry for Economic Affairs and Energy [8]. IntelMOD investigates how an online Intelligent Modernization Platform (IMP) supported by AI can best support decision-making for planning processes of climate-friendly modernization measures in existing residential buildings [8]. The Hamburg University of Applied Sciences is responsible for exploring innovative interactions, as well as developing the AIVA part of the platform. IntelMOD builds in part on its predecessor KIMM [7], which implemented the *functional cost splitting* method developed by David [9], a method for separating modernization costs from maintenance costs as a transparent basis for communication between landlords and tenants. Since this method still requires a lot of technical and domain knowledge, the AIVA is intended to assist the user in making informed choices.

Two AIVAs were developed as part of the IntelMOD project. Application A (Figure 5.1) is the guided assistant, and Application B (Figure 5.2) is the basic chat. Both of the applications are described in detail in Table 5.1.

The phases of the guided assistant are as follows:

Phase 1 Introduction and prompt for address

Phase 2 Collect building information

Phase 3 Collect current functional building condition

Phase 4 Rank aesthetics, environmental protection, and cost

Phase 5 Provide the building modernization recommendation

Phase 6 Display the result calculated by the FK-API with an optional follow-up Q&A

Table 5.1: Comparison of the two applications

	Application A: Guided Assistant	**Application B: Basic Chat**
Interaction	System-led, six fixed phases	User-led question answering
Task	Collect the 20 inputs of the cost-splitting form via dialogue, recommend a measure, present the cost split	Explain terms, procedures, and case law while the user fills in the form in the *Expertenmodus* (split screen)
Output	Short questions, a server-built top-3 recommendation, a result table	Explanatory answers with source references
Deterministic parts	Phase control, value validation, recommendation scoring, cost calculation (FK-API), fixed texts	Post-processing of the output only
LLM calls per user turn	1–4 (classification, extraction, phrasing)	1
Retrieval	Only in the follow-up phase (Phase 6) on the user's question	On every user message
User interface	Chat with progress bar and phase sidebar	Narrow chat window

5.2 Requirements

In addition to the design goals, there are a number of requirements the AIVAs have to fulfill. All of the functional (F) and non-functional (NF) requirements regarding the AIVA collected during the work in the IntelMOD project are displayed in Table 5.2.

5.3 Design Constraints

Two regulatory frameworks act as constraints on the system design. First, the GDPR [24] applies as users describe their individual housing situation and may include personal data such as rent, address, or income. Second, the newly introduced EU AI Act [147, Art. 50(1)] imposes transparency obligations on AI systems that interact directly with persons. The



Figure 5.1: Overview of Application A: Guided Assistant (German user interface)

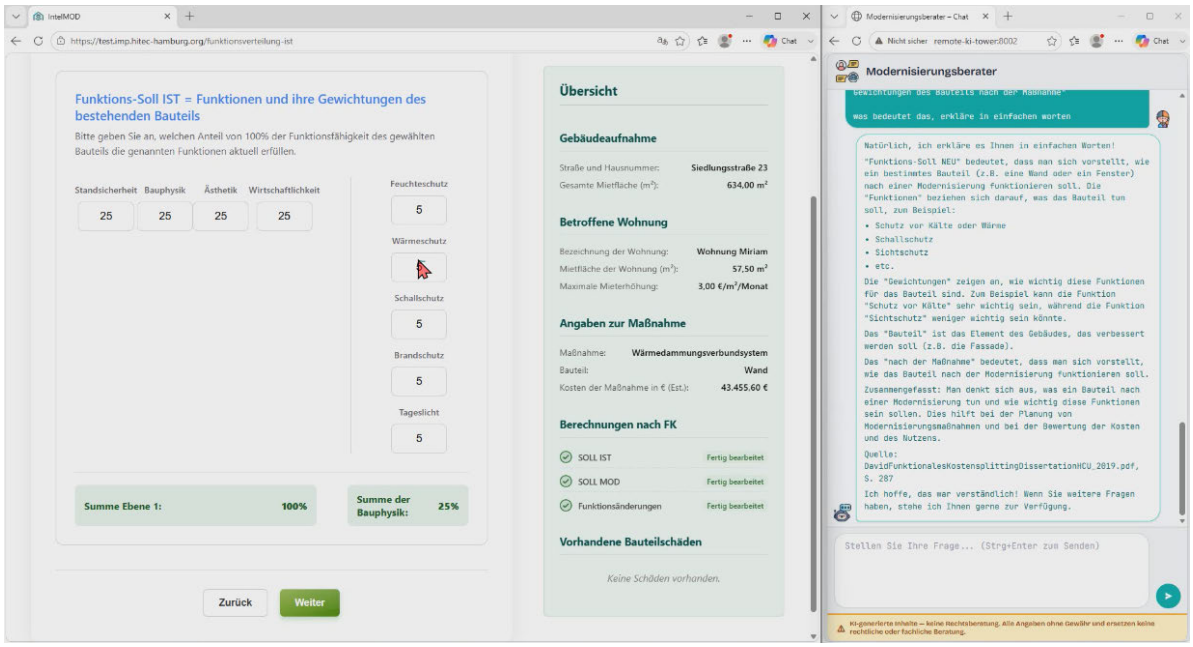

Figure 5.2: Overview of Application B: Basic Chat (German user interface)

Table 5.2: Functional and non-functional requirements and their implementation

ID	Requirement	Type	Implemented by
FR1	Collect the 20 form inputs in German dialogue, several values per message	F	LLM extraction to JSON (`get_value_extraction_prompt`) plus a position-based numeric fallback
FR2	Distinguish answers from follow-up questions and answer those questions	F	FRAGE/ANTWORT classification (heuristics first, then LLM) plus a clarification prompt
FR3	Allow users without building documents to continue	F	Default values for the current-condition ratings (`IST_STANDARD_VALUES`)
FR4	Recommend suitable measures based on user priorities, with a transparent rule	F	Deterministic weighted scoring (Equation (5.1))
FR5	Calculate the cost split and rent increase according to § 559 BGB	F	External FK-API; the LLM only presents the result
FR6	Answer free questions using the knowledge base and cite the sources	F	RAG (B: every turn; A: Phase 6). The prompt asks for author, year, and page
NFR1	The dialogue must never get stuck	NF	Max. 4 attempts per field, then a default value; deterministic fallback texts; `/fallback` command
NFR2	Identical conditions for both model sizes	NF	Same code, prompts, parameters, RAG index and UI. Only `LLM_BACKEND`/`LLM_MODEL` differ (Docker Compose profiles)
NFR3	Disclose the use of AI and the limits of the advice	NF	Permanent banner “KI-generierte Inhalte – keine Rechtsberatung …” in both UIs
NFR4	Minimize data transfer to third parties	NF	Local 8B, local embeddings and vector store (the 70B is an explicit exception, see Section 5.3)
NFR5	Log interactions for analysis	NF	One JSON file per session, written atomically on every turn (`conversation_logger.py`)
NFR6	Formal German, plain text without markdown	NF	Prompt rule plus `strip_markdown()`

AIVA developed in IntelMOD therefore requires compliance with the GDPR and the EU AI Act, as well as explainability of any calculations done by the system.

The GDPR [24] in particular favors models that can be hosted on-premises so that no user data is passed to third parties and used for data retention or training, whereas the transparency obligation of the AI Act [147] is met by disclosing the use of AI in the user interface (NFR3). The Apertus models [146] meet the conditions, as weights, training data, and code are openly available. The training data consists only of publicly available web data. Personally identifiable information was anonymized, and post-training data was filtered by license [146]. Because the IntelMOD project with its IMP is developed for the German market, a model capable of the German language is crucial to ensure compatibility. In the case of Apertus, German is the third-largest language in its multilingual pre-training data [146].

Since running a 70B model locally requires considerably more hardware than an 8B model, surpassing most consumer GPU capabilities, and inference energy consumption grows with model size [23], the question arises whether a small, locally hosted language model is sufficient from a user's perspective. This is the motivation for the comparison of an 8B and a 70B model in this thesis. In this study, the 70B model was accessed through the Hugging Face Inference Providers router [148] and served by Public AI [149]. As this took place within a controlled study environment, it was deemed acceptable to use an external provider. For production use, a local solution should be preferred.

5.4 System Architecture

Both applications share one codebase (Table 5.3), with the two model configurations differing only in the model and its hosting (Table 5.4). Figures 5.3 and 5.4 give an overview of how the components of the two applications are connected, distinguishing between LLM-based and deterministic components.

Application A follows a hybrid design in which the LLM is confined to language tasks, whereas all decisions and numbers are produced deterministically. The LLM is used for four tasks: (1) classifying whether a user message answers the current question or asks a follow-up question, (2) extracting field values from free text into JSON, (3) phrasing the questions and transitions, and (4) answering free follow-up questions after the calculation. The dialogue state, i.e., the current phase, the collected values, and the transitions between phases, is managed by the server. Measures are ranked with a transparent scoring rule: the user's ranking of cost, environmental protection, and aesthetics is converted into weights $w \in \{3, 2, 1\}$, and each of the six possible modernization measures m receives the score

$$s_m = w_{\text{cost}} \left(1 - \frac{p_m}{\max_k p_k}\right) + w_{\text{eco}}\, e_m + w_{\text{aes}}\, \frac{a_m}{100}, \tag{5.1}$$

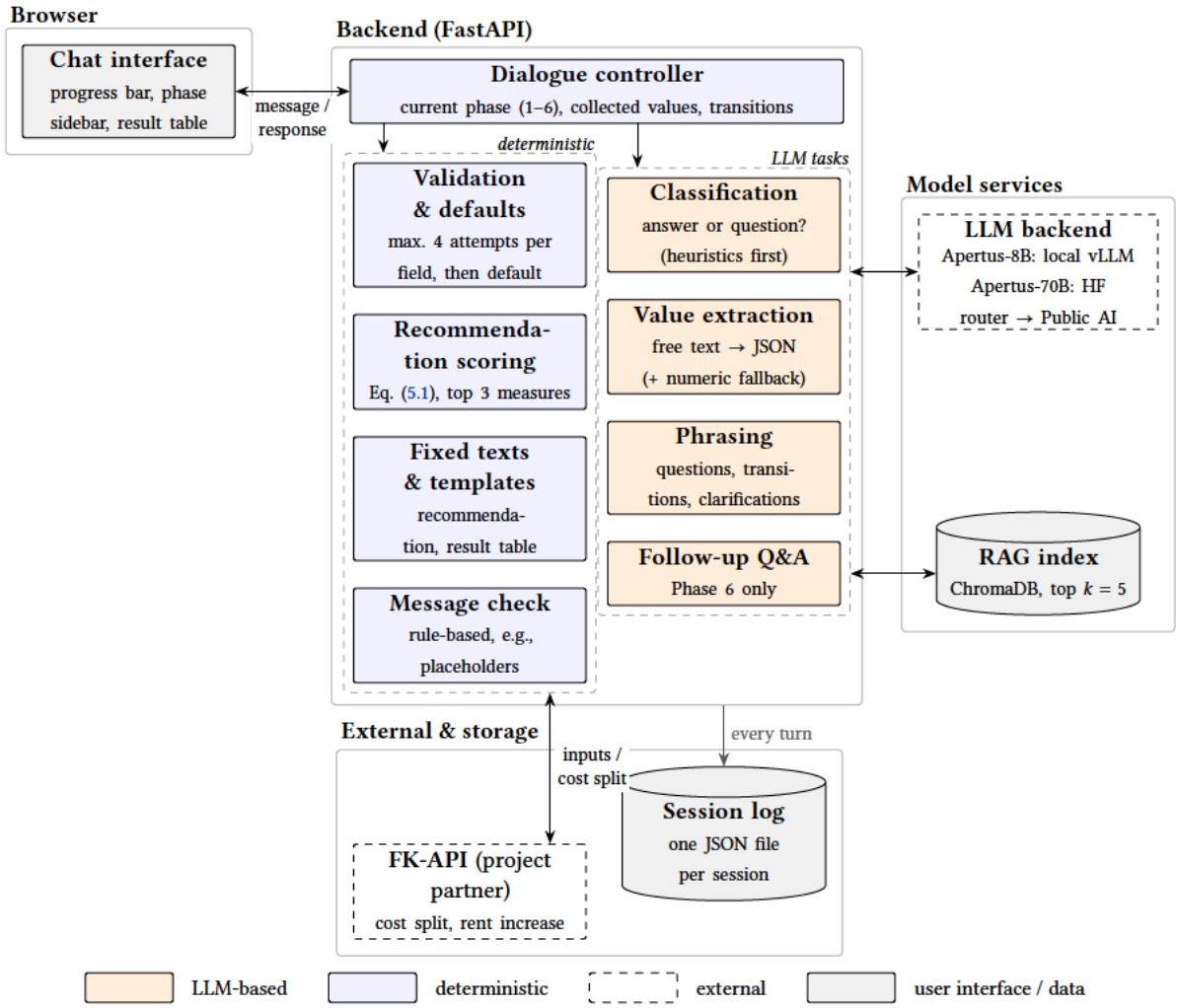

Figure 5.3: Architecture of Application A (Guided Assistant). The server-side dialogue controller confines the LLM to language tasks (one to four calls per user turn), whereas validation, the recommendation, the cost calculation via the FK-API, and the result texts are deterministic.

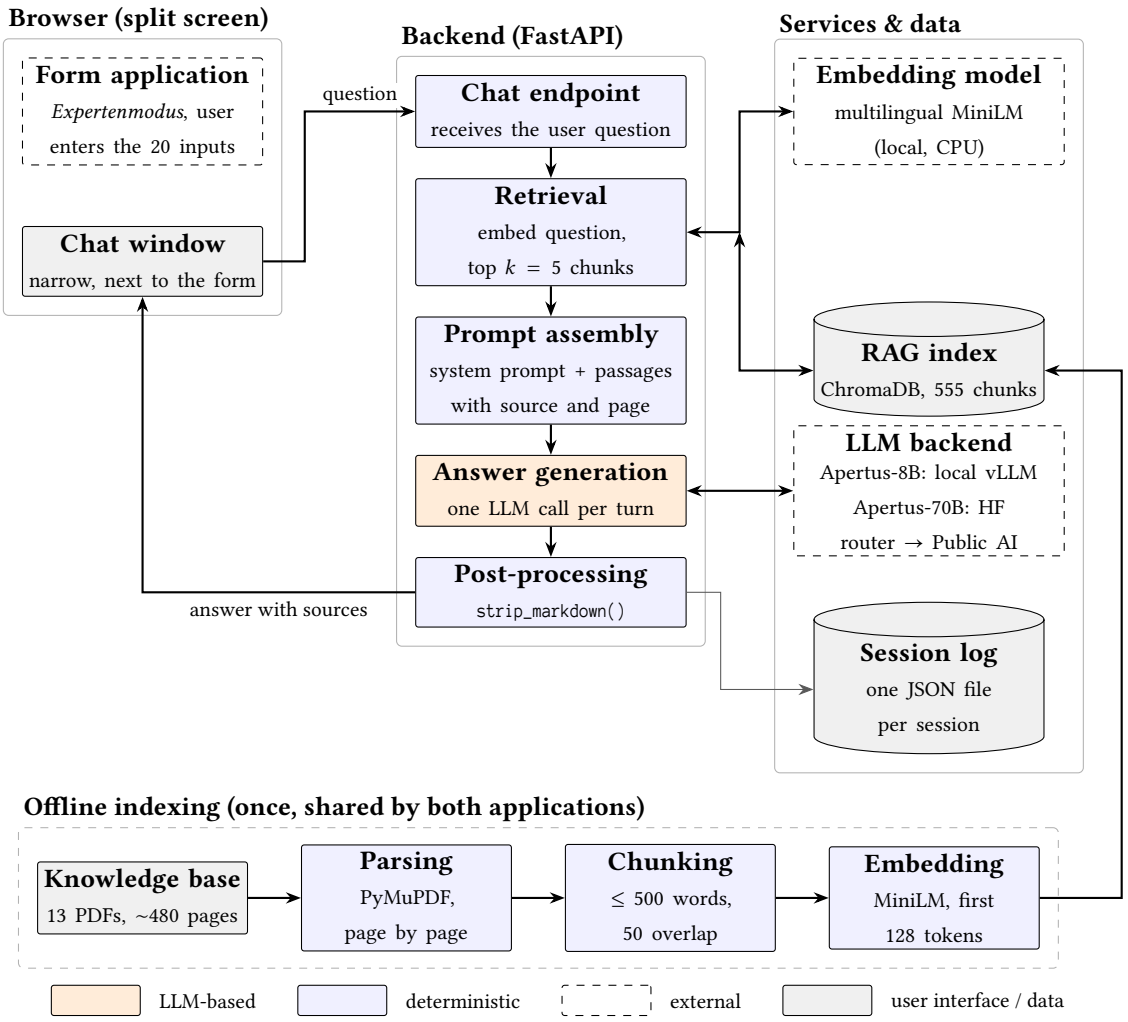

Figure 5.4: Architecture of Application B (Basic Chat). The user fills in the form application and asks the assistant questions in a chat window next to it. Every question triggers retrieval from the RAG index, which is built once from the knowledge base, followed by a single LLM call. The form application itself is not part of the AIVA and is not described further.

Table 5.3: Technology stack of the prototype

Layer	Technology
Frontend	Plain HTML/CSS/JavaScript, `marked` for markdown rendering of the result table
Backend	Python 3.12, FastAPI, uvicorn
LLM serving (local)	vLLM (`vllm/vllm-openai:0.23` image), bfloat16, `--max-model-len 64K`
LLM serving (remote)	Hugging Face Inference Providers router, provider `publicai`
Embeddings	`sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` (local, CPU)
Vector store	ChromaDB `PersistentClient`, HNSW index, cosine distance
PDF parsing	PyMuPDF
Deployment	Docker Compose, NVIDIA Container Toolkit

Table 5.4: Configuration of the small and medium language models

	Small	Medium
Model	`swiss-ai/Apertus-8B-Instruct-2509`	`swiss-ai/Apertus-70B-Instruct-2509`
Hosting	Local, vLLM, NVIDIA RTX 5090	Remote, Hugging Face router → Public AI
Precision	bfloat16 (no quantization)	bfloat16 (no quantization)
Context window	64K tokens configured	Provider default
Temperature / top-p / frequency penalty	0.8 / 0.9 / 0.15	identical
Max. new tokens	B: 2,048; A: 10 (classification), 50 (ranking), 400 (extraction), 512 (phrasing), 800 (Phase 6)	identical
Streaming	No (full response is returned at once; typing indicator in the UI)	identical

where p_m is the price per square meter, $e_m \in [0, 1]$ the ecological sustainability score, and $a_m \in [0, 100]$ the aesthetic function score of measure m, so that each term lies between 0 and 1 before weighting. The top three measures and the reasons shown for them are inserted into the response by the server, and the LLM only writes a one-sentence introduction. The calculation of the functional cost splitting is done via the FK-API, an external API provided by a project partner that takes all of the necessary values and returns a deterministic cost calculation. This is done to ensure that the calculation is not susceptible to the AIVA behavior and that calculations remain explainable.

Application B follows a much simpler design with the AIVA using only a single LLM call per user turn with the retrieved passages as additional context. Because this application was developed to be used side by side with the more complex form application, it serves only the question-and-answer use case. As this thesis focuses only on the AIVA side, the form application is not described in more detail.

Prompt Design Both applications use German system prompts that define the assistant's role as a friendly expert advisor, define the formal form of address, and list rules, as well as a glossary of domain abbreviations. The full prompts are given in Appendix A.7. No dedicated guardrail model such as Llama Guard [150] or a guardrail framework such as NeMo Guardrails [151] was used, since each of their checks requires additional LLM calls and thus adds latency [151]. Instead, Application A validates every generated message with a rule-based check to catch accidental placeholder names or other mistakes. The prompts and checks were developed iteratively in a pilot phase in which all LLM requests and responses were logged and evaluated.

5.5 Retrieval-Augmented Generation

In order to enhance the answers with domain knowledge, both applications use retrieval-augmented generation [116] in its basic "Retrieve-Read" form [117].

The knowledge base consists of a total of 13 German PDF documents with about 480 pages. This includes eight rulings of the Bundesgerichtshof (Federal Court of Justice) (BGH) on tenancy law and rent increases, two law journal articles, an overview of BGH cases on tenancy law, a commentary excerpt, and the dissertation by David [9] on functional cost splitting (Table 5.5). The documents were chosen as part of the KIMM and IntelMOD projects [7], [8].

PyMuPDF reads the documents page by page, so that every passage keeps track of its page number. The pages are then split into chunks of a maximum of 500 words with an overlap of 50 words so that most pages form a single chunk, yielding 555 chunks (with a

median of 251 words). As larger chunks carry more context but also more noise, and smaller chunks less context [117], page-sized chunks were chosen so that every retrieved passage can be directly cited with a page number. The chunks are embedded with the multilingual sentence-transformer *paraphrase-multilingual-MiniLM-L12-v2* [152], which runs locally with a persistent ChromaDB collection. The embedding model truncates its input to 128 tokens, so only roughly the first 80–100 words of each chunk enter its vector.

The retrieval step is dense, meaning that questions and passages are compared as vectors, capturing paraphrases that keyword search would potentially miss [153]. This matters especially in technical domains where lay users describe terms and concepts in everyday language. For each user message, the five most similar chunks ($k = 5$) are retrieved with their source and page before being appended to the system prompt together with the instruction to correctly cite the source. Figure 5.5 shows an example retrieval process.

Table 5.5: Composition of the RAG knowledge base

Document type	**Files**	**Pages**	**Words (pdftotext)**
BGH rulings on modernization rent increases (2007–2023)	8	101	~15,700
Journal articles (Börstinghaus, NZM 2021; Schindler, NZM 2021)	2	43	~21,800
beck-online overview "Die aktuelle Rechtsprechung des BGH zum Mietrecht"	1	34	~16,700
Commentary excerpt (MüKo BGB 2023, Artz)	1	2	~760
David [9], dissertation on functional cost splitting	1	300	~65,800
Total	**13**	**480**	**~121,000**

5.6 Technical Validation

To assess the technical properties of the language models' behavior, 20 typical questions (form terms, concepts of functional cost splitting, and tenancy law) were sent to both the 8B and the 70B LLMs three times through the application pipeline. Responses that the inference provider returned from its cache were excluded in order to ensure better comparability, leaving 60 unique generations from the 8B model and 29 from the 70B model (Table 5.6).

The locally hosted 8B model answered after a median of 3.9 s (Interquartile Range (IQR) 3.1–5.2 s), and the remotely hosted 70B model after 32.4 s (IQR 28.0–41.9 s). The difference results mainly from the generation speed of the two serving set-ups (94 vs. 17 tokens per

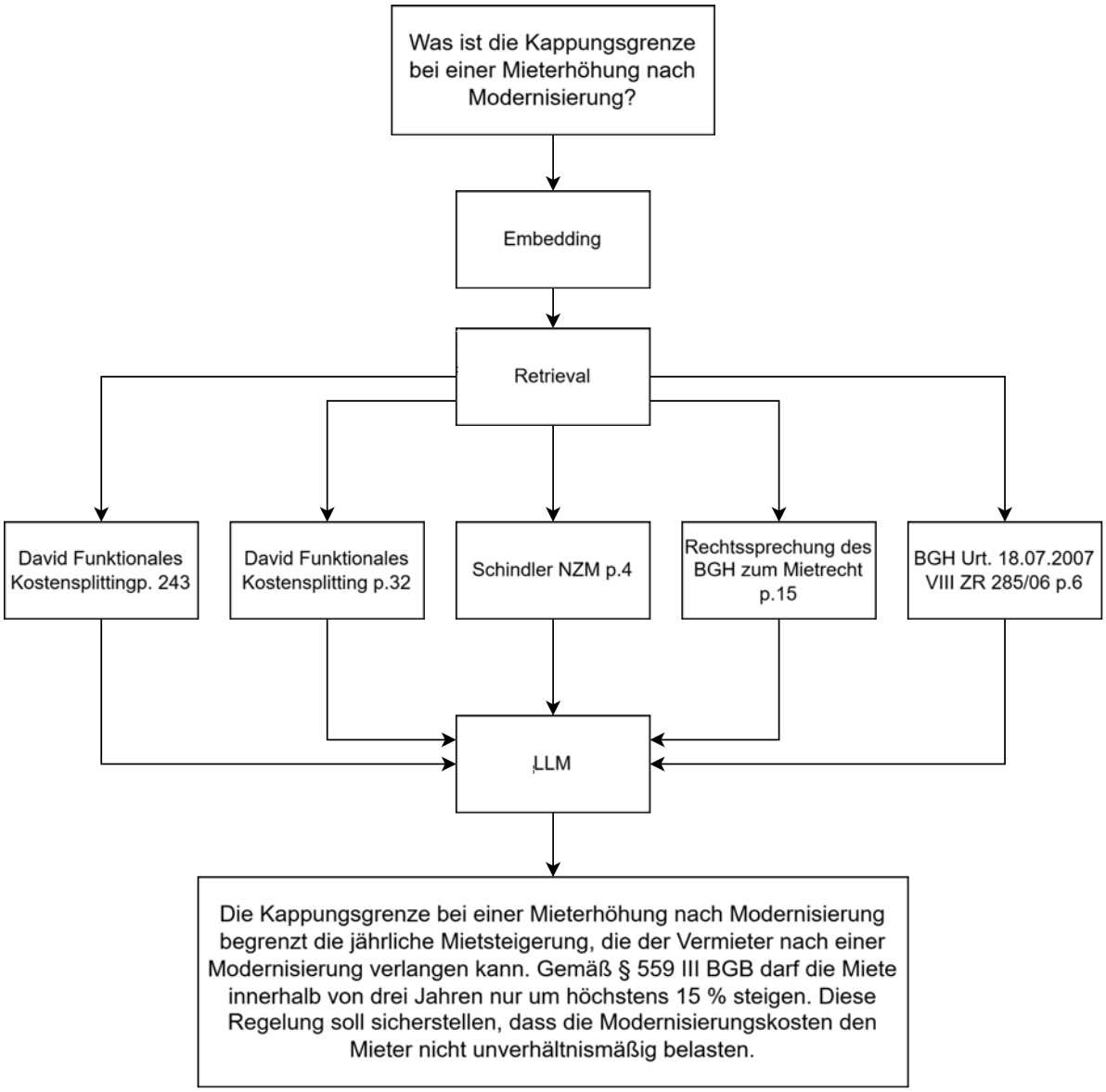

Figure 5.5: Example process of retrieval for a technical question using the RAG component

second) and only partly from answer length: the 70B model wrote longer answers (on average 2,223 vs. 1,562 characters), and none of the answers reached the limit of 2,048 tokens. This mirrors the response lengths observed in the study (Section 7.1.2). Because responses were not streamed, these latencies correspond to the time a user waits for an answer. Most answers contained a reference to a source (72 % for the 8B and 86 % for the 70B model), but only 32 % and 41 % referred to a document that had actually been retrieved; the references mostly reproduced file names, and some named sources that were not part of the retrieved context. Retrieval itself was dominated by the dissertation on functional cost splitting, which provided 76 of the 100 retrieved passages, whereas the BGH rulings were rarely retrieved. During the study, the FK-API failed to respond in three sessions and in 6 of the 24 Application A sessions (two with the 8B and four with the 70B model), the *fallback* command was used. The fallback command is a manual override of the users' inputs issued by the researchers to be used when the system is stuck and does not progress the conversation. It replaces the collected inputs with a predefined sample building and proceeds directly to the calculation.

Table 5.6: Comparison of the 8B and 70B models in the technical pilot

	8B (local, vLLM)	**70B (remote, Public AI)**
Requests / unique generations	60 / 60	60 / 29
Errors	0	4 × HTTP 504 gateway timeout after ~61 s (= 12 % of the 33 uncached attempts)
Latency per answer, *Mdn* [IQR]	3.9 s [3.1–5.2]	32.4 s [28.0–41.9]
Latency range	2.0–8.2 s	14.2–54.1 s (+ 4 timeouts)
Generation speed, *Mdn*	94 tokens/s	17 tokens/s
Output tokens, *Mdn*	363	550
Characters, *M* (*SD*)	1,562 (580)	2,223 (789)
Answers cut off at 2,048 tokens	0 %	0 %
Any source reference	72 %	86 %
Reference to a retrieved document	32 %	41 %
Markdown in raw output (removed before display)	48 %	59 %

6 Methodology

This chapter describes the user study conducted to test the hypotheses and answer the research questions. It covers the study design (Section 6.1), the measures (Section 6.2), the study procedure (Section 6.3), the recruitment of participants and ethical considerations (Section 6.4), and the analysis strategy (Section 6.5).

6.1 Study Design

The study followed a 2×2 factorial design with the application (A vs. B) and the language model size (8B vs. 70B) as the factors. Every participant interacted with each application with either the same model or both models, as the model size was assigned independently for the two tasks. Crossing the two possible orders of the applications with the four model size combinations results in eight sequences, displayed in Table 6.1. This allows for each application × model combination to be observed 12 times with a balanced application order, and each model appears equally often in the first and second positions.

Using only four sequences would have been sufficient to balance the order of the applications but not each model size. Using all eight sequences ensures that every application × model combination appears equally often in each position. This approach balances position and first-order carry-over effects [154].

The sample size of $N = 24$ was chosen as a multiple of the eight sequences and with regard to the resources available in the research project. It is above the median of 15 participants in in-person studies at CHI and the mean of 17 in within-subjects experiments [155]. However, local standards are no substitute for a power analysis: Caine [155] notes that many studies published at CHI are underpowered even for large effects. A sensitivity analysis in G*Power 3.1 [156], [157] (linear multiple regression, R^2 increase; $\alpha = .05$, $1 - \beta = .80$) showed that, with $n = 24$ observations per application, both a single predictor and the interaction term in a model with three predictors can only be detected reliably if $f^2 \geq .36$. This corresponds to a large effect [158]. The study is therefore powered to detect strong relationships such as those postulated by the core UTAUT paths, but not small or medium effects; non-significant results should accordingly not be interpreted as evidence for the absence of an effect.

Table 6.1: Task sequences and assigned participants

Sequence ID	Task 1	Task 2	Participants
P1	A + 8B	B + 8B	TP1, TP9, TP17
P2	A + 8B	B + 70B	TP2, TP10, TP18
P3	A + 70B	B + 8B	TP3, TP11, TP19
P4	A + 70B	B + 70B	TP4, TP12, TP20
P5	B + 8B	A + 8B	TP5, TP13, TP21
P6	B + 8B	A + 70B	TP6, TP14, TP22
P7	B + 70B	A + 8B	TP7, TP15, TP23
P8	B + 70B	A + 70B	TP8, TP16, TP24

6.2 Measures

Demographic questions regarding age, gender, German proficiency, education, technology affinity, chatbot usage, and tenancy law knowledge were asked. The complete demographic questionnaire is shown in Appendix A.1.

Perceived usability was measured with the mixed-tone German translation of the SUS (Appendix A.3) by Rummel [93], with ten items rated on a 5-point Likert scale and scored from 0 to 100 [89]. This version has been psychometrically validated and recommended among researchers [92], [94]. Scores are interpreted using the curved grading scale by Sauro and Lewis [95], in which the average score of 68 corresponds to a grade of C.

Technology acceptance was measured with 17 items of the UTAUT questionnaire [11] covering PE (4 items), EE (4), SI (2), FC (4), and BI (3), each rated on a 7-point Likert scale. The two organizational SI items (senior management and organization) were omitted because no organizational context existed. The remaining items kept the original wording and therefore partly refer to a work context (e.g., “in meiner Arbeit”, “Gehaltserhöhung”). The negatively worded FC item (“nicht kompatibel”) was reverse-coded. The German translation (Appendix A.5) was based on the validated German translation of the UTAUT2 questionnaire by Harborth and Pape [159], using the original UTAUT wording where the two versions differ.

For this thesis, the dimensional valence–arousal model is used because it covers the full affective state rather than just a small set of prototypical emotions [65] and valence is the aspect of facial behavior that is most reliably perceived [64]. Additionally, earlier acceptance research focused on single affective constructs such as anxiety or enjoyment, whereas using valence and arousal allows affect to be modeled more completely [20], [126]. Thus, facial expressions were recorded with a webcam and analyzed with the AFFDEX module [67] integrated in the iMotions software [68]. AFFDEX detects facial landmarks, classifies facial

action units, and derives emotion metrics from them for every video frame [67]. Displayed valence was captured by the *Adaptive Valence* metric (−100 to 100) and displayed arousal by *Adaptive Engagement* (0 to 100), averaged per session [69]. In addition to the dimensional metrics, AFFDEX reports the six basic emotions [53] anger, disgust, fear, joy, sadness, and surprise with the addition of contempt (0 to 100) [67], [69], which were likewise averaged per session to examine whether a categorical view of affect explains acceptance better than valence and arousal.

6.3 Study Procedure

After being welcomed and provided with a detailed explanation of the study in order to give their informed consent, the participants completed the demographic questionnaire. They then received the first task scenario in which they were given a description of a building in need of modernization and tasked with using the first application for as long as they needed to plan the modernization measure while their faces were recorded. Directly afterwards, they rated the application with the SUS and the UTAUT questionnaire. The same procedure was followed for the second application. The session ended with a short debrief and took about 60 minutes in total.

The study was run in iMotions [68], which synchronizes the webcam recording, the AFFDEX analysis, and the questionnaires in one study environment. Using iMotions in this study setting allows for easy organization of participant data as well as efficient collection of the FER data stream and the questionnaires. Figure 6.1 shows the general view of a study inside the iMotions study software, while Figure 6.2 shows the study sequence builder function.

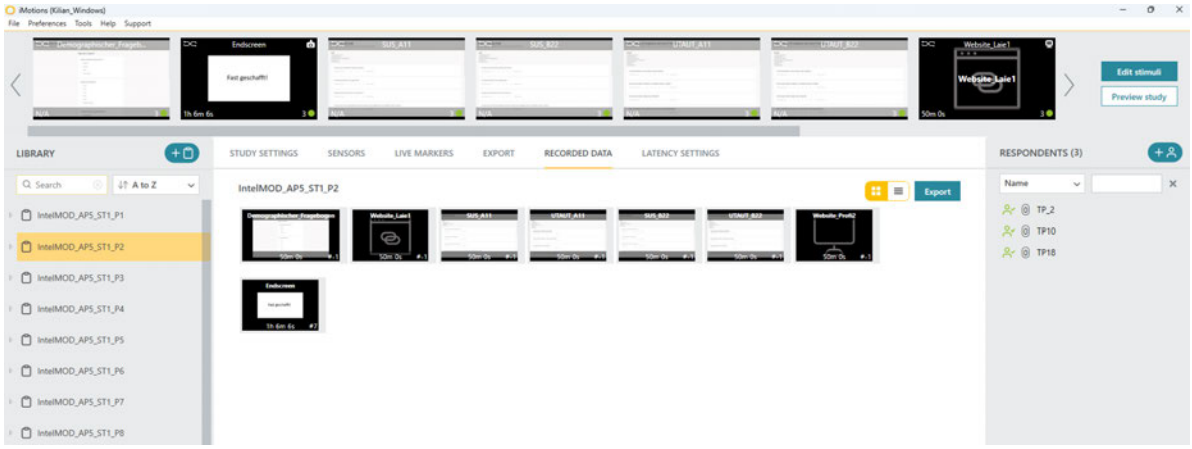

Figure 6.1: Overview of the iMotions software study view

The exact order of actions is displayed in Figure 6.3.

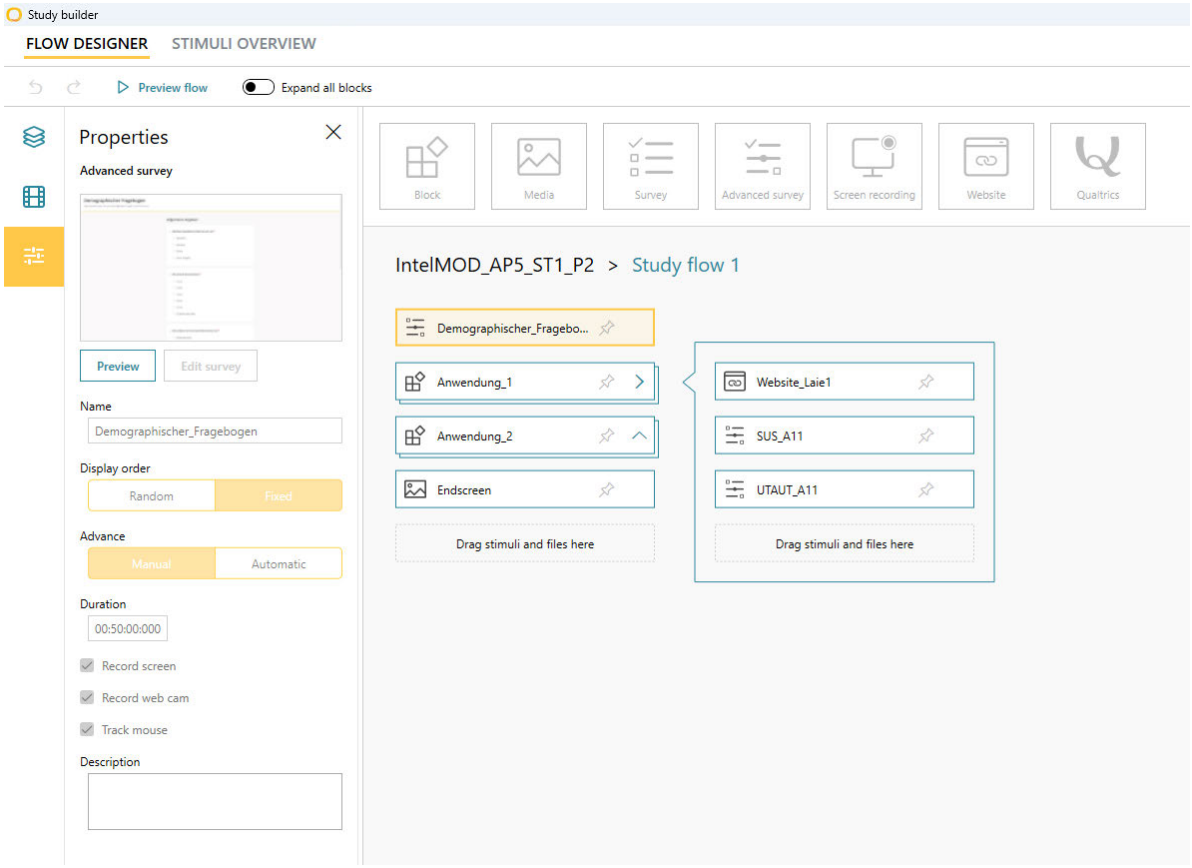

Figure 6.2: Overview of the iMotions study builder

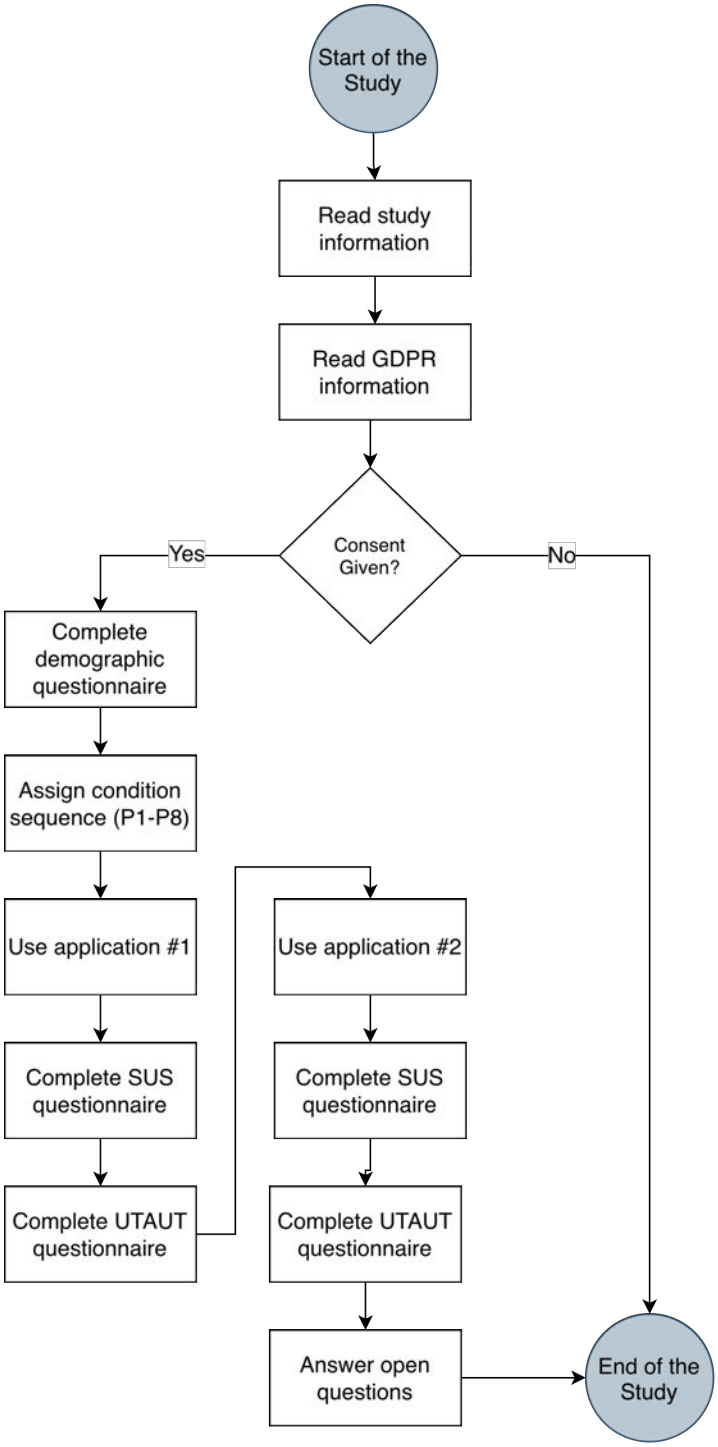

Figure 6.3: Order of actions for the study

6.4 Recruiting and Ethical Considerations

Participants were recruited through the Artificial Intelligence Center Hamburg (ARIC) [160] newsletter, a networking event, personal contacts, and the university context. Participants had to be at least 18 years old and be proficient in German. They were compensated with 15 €. The study was approved by the ethics committee of the Hamburg University of Applied Sciences (Appendix A.6). Participants gave written informed consent, including consent to video recording of their face, and were informed that they could withdraw at any time without giving reasons. Video data were stored encrypted via the VeraCrypt software [161] and processed only to extract the facial expression metrics. All other data were processed in pseudonymized form.

6.5 Analysis Strategy

All analyses were conducted in Python 3.14 with the following packages:

- pandas 3.0.3 [162], [163]
- NumPy 2.5.1 [164], [165]
- SciPy 1.18.0 [166]
- statsmodels 0.14.6 [167]
- Matplotlib 3.11.0 [168]
- PyArrow 25.0.1 [169]

Internal consistency was assessed with Cronbach's α [104].

The effects of application, model size, and their interaction on usage, SUS, valence, arousal, and the UTAUT constructs were estimated with linear mixed-effects models with a random intercept per participant. The sequence (starting with Application A or B) was included as a fixed covariate, and all factors were sum-coded (±0.5). Coefficients were estimated with restricted maximum likelihood (REML). Each fixed effect was tested with a likelihood-ratio test against the full model without this term, refitted with maximum likelihood, and the p values were corrected for the four terms per outcome using the Holm method [170]. Normality of residuals was checked with the Shapiro–Wilk test [171] and homogeneity of variances with Levene's test [172]. As each participant contributes one observation per application, the hypotheses were tested separately for each application ($n = 24$) so that observations are independent. H1a and H1b were tested with simple linear regression and H2a–H2c with hierarchical moderated regressions. Model 1 contained displayed valence and arousal

(grand-mean centered within the application), and Model 2 added their interaction. Influential cases were identified with Cook's distance ($D > 4/n$) [173]. To test H1c, the correlations of perceived usability with effort expectancy and with performance expectancy were compared within each application. Because both correlations share the SUS and stem from the same participants, the test for dependent overlapping correlations by Meng et al. [174] was used, with a 95 % confidence interval for the difference after Zou [175]. As robustness checks, the comparison was repeated without the influential case of the H1a models and with Spearman rank correlations.

Spearman correlations were computed between usage metrics, demographics, SUS, and UTAUT ratings. The resulting p values were corrected with the Benjamini–Hochberg procedure [176].

As a robustness check, all hypothesis models were also estimated on the 48 sessions of both applications together with a random intercept per participant and application, model size, and session position as fixed covariates (Table A.8 in the appendix).

7 Results

This chapter reports the results of the user study. It first describes the sample and how the applications were used (Section 7.1), then presents the descriptive results and the effects of application and model size on perceived usability, displayed affect, and the UTAUT constructs (Section 7.2). The hypothesis tests follow in Section 7.3, and exploratory findings are in Section 7.4.

7.1 Sample Description

This section describes the participants based on the demographic questionnaire and summarizes how long and how intensively they used the two applications.

7.1.1 Demographic Questionnaire

The sample consisted of $N = 24$ participants (14 male and 10 female, no one identified as non-binary), with a median age category of 25–34. Participants were highly educated (23 of 24 participants held at least the Abitur, the German equivalent of a high school diploma) and reported a high affinity for digital applications, with frequent use of chatbots/AI assistants (71 % at least weekly). Self-rated knowledge regarding tenancy laws was low throughout the sample, with none of the participants selecting “high” or “very high”. Appendix A.1 shows all of the demographic questionnaire results.

7.1.2 Application Usage

Participants spent, on average, 12.9 minutes in an application ($SD = 5.8$; Table 7.1, Figure 7.1): 10.3 minutes in Application A and 15.6 minutes in Application B. Application A produced about twice as many LLM responses per session ($M = 5.7$ vs. 2.8), but these were much shorter (463 vs. 1,921 characters on average). In Application B, the medium model wrote markedly longer responses than the small model (2,512 vs. 1,389 characters), whereas in Application A the two models hardly differed (498 vs. 426 characters).

Table 7.1: Usage by application × language model, *M* (*SD*)

Application	Model	Time (min)	Responses	Characters	n_{char}
Application A	8B	9.5 (5.3)	6.1 (2.0)	426 (113)	11
Application A	70B	11.0 (3.5)	5.4 (1.9)	498 (164)	12
Application B	8B	15.9 (6.5)	3.7 (2.8)	1,389 (660)	10
Application B	70B	15.3 (5.5)	1.9 (1.8)	2,512 (1,221)	9
Overall		12.9 (5.8)	4.3 (2.7)	1,123 (1,041)	42

Note. Time = StartSlide to EndSlide of the task. Responses = LLM replies per session. Characters = mean characters per LLM reply (greeting, error messages and fixed templates excluded); sessions without any LLM reply have no value (n_{char}).

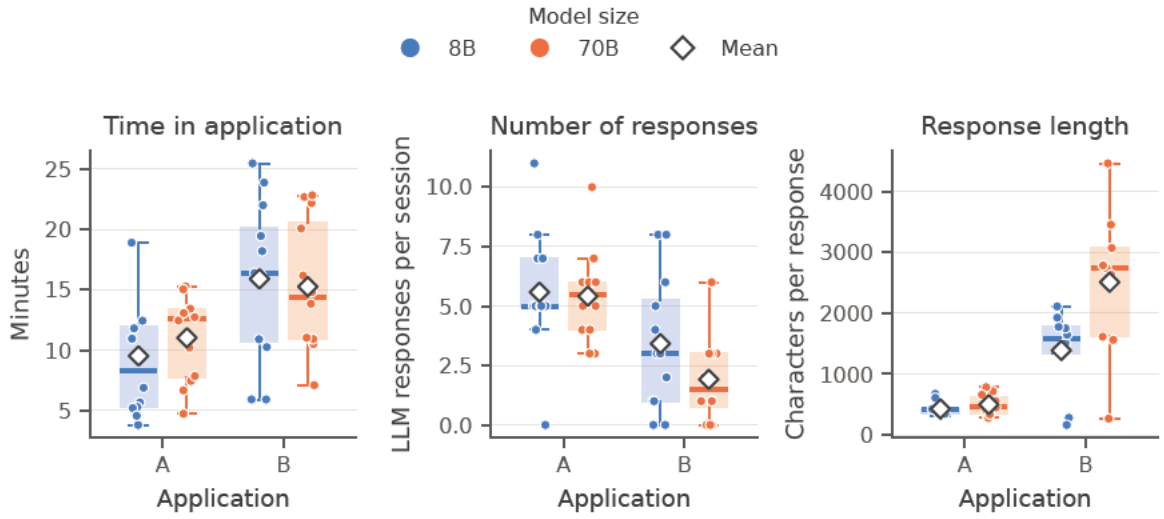

Figure 7.1: Time in application, number of LLM responses and mean response length by application and model size. Boxes show the interquartile range and median, diamonds the mean, and dots the individual sessions.

Participants spent 5.3 minutes longer in Application B than in Application A (95 % CI [2.3, 8.3], $d = 0.91$, $\chi^2(1) = 11.76$, $p < .001$, Holm-corrected $p = .002$; Table 7.2), whereas the model size had no effect on the time in the application ($\chi^2(1) = 0.09$, $p = .767$). The response length was affected by all four terms. Responses in Application B were 1,513 characters longer than in Application A (95 % CI [1,134, 1,892], $\chi^2(1) = 41.09$, $p < .001$), and the small model wrote 578 fewer characters than the medium model (95 % CI [−959, −197], $\chi^2(1) = 9.12$, $p = .003$, Holm-corrected $p = .008$). The significant interaction ($\chi^2(1) = 7.70$, $p = .006$, Holm-corrected $p = .011$) shows that this difference arose mainly in Application B. The sequence also had an effect ($\chi^2(1) = 5.53$, $p = .019$). Because the residuals of this model were neither normally distributed ($W = 0.93$, $p = .009$) nor homogeneous ($F = 5.43$, $p = .003$), the p values for response length are only approximate; the descriptive pattern, however, is clear.

Table 7.2: Likelihood-ratio tests for usage ($\chi^2(1)$ and Holm-corrected p)

	Model size		Application		Interaction		Sequence	
Outcome	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}
Time in application (min)	0.09	1.000	11.76	.002	0.54	1.000	0.92	1.000
Characters per LLM response	9.12	.008	41.09	$< .001$	7.70	.011	5.53	.019

7.2 Descriptive and Inferential Statistics

This section reports perceived usability, displayed valence and arousal, and the UTAUT ratings for each combination of application and model size, together with the likelihood-ratio tests of the linear mixed-effects models described in Section 6.5. The main effects of application and model size in these models are the tests of H3a–H4c, which are evaluated in Section 7.3.4.

7.2.1 SUS

Table 7.3 shows the mean SUS scores, the standard deviations, and the corresponding letter grades [95]. Application A with the small model received the highest rating (grade A), whereas Application B with the medium model received the lowest (grade D). Averaged over all 48 ratings, the SUS score was 73.7 (grade B−), slightly above the commonly cited average of 68 or the newer mean score of 70.8 [19], [177]. Figure 7.2 shows that the ratings for Application B were considerably more dispersed: the three lowest scores (2.5, 12.5, and 20.0) all came from Application B with the medium model.

Application A received significantly higher SUS scores than Application B, by 14.5 points (95 % CI [6.0, 23.0], $d = 0.64$, $\chi^2(1) = 10.04$, $p = .002$, Holm-corrected $p = .006$; Table 7.4).

Seventeen of the 24 participants rated Application A higher, and a Wilcoxon signed-rank test confirmed the difference ($p = .004$). The medium model was rated 7.9 points lower than the small model (95 % CI $[-18.4, 2.7]$, $d = -0.35$), but this difference was not significant ($\chi^2(1) = 2.35$, $p = .125$, Holm-corrected $p = .375$), and neither the interaction nor the sequence had an effect ($\chi^2(1) \leq 0.05$, $p \geq .829$). The residuals of the SUS model were normally distributed ($W = 0.97$, $p = .232$), and the variances were homogeneous ($F = 1.95$, $p = .135$).

Table 7.3: SUS scores by application × language model

Application	Model	Mean	SD	Grade
Application A	8B	83.5	18.3	A
Application A	70B	78.3	12.3	B+
Application B	8B	73.1	20.5	B−
Application B	70B	59.8	31.1	D
Overall		73.7	22.8	B−

Note. $n = 12$ sessions per cell. Grades are from the curved grading scale of Sauro and Lewis [95].

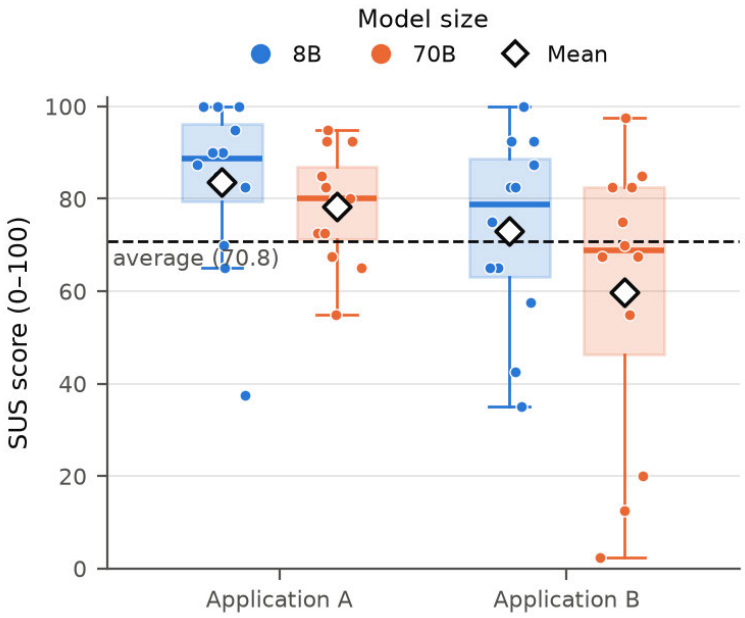

Figure 7.2: SUS scores by application and model size. Boxes show the interquartile range and median, diamonds the mean, and dots the individual sessions ($n = 12$ per box). The dashed line marks the average SUS score of 70.8 [177].

Table 7.4: Likelihood-ratio tests for the SUS ($\chi^2(1)$ and Holm-corrected p)

	Model size		Application		Interaction		Sequence	
Outcome	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}
SUS	2.35	.375	10.04	.006	0.05	1.000	0.04	1.000

7.2.2 Valence & Arousal

A face was detected in 82 % of the expected video frames on average (minimum 56 %). Displayed valence was close to neutral and slightly positive on average ($M = 6.67$, $SD = 10.78$, $Mdn = 2.39$ on a scale from −100 to 100), with a few sessions showing clearly positive expressions (Table 7.5, Figure 7.3). Arousal averaged $M = 23.92$ ($SD = 15.07$) on a scale from 0 to 100. Both measures varied far more between participants than between conditions.

Table 7.5: Displayed valence and arousal by application × language model, M (SD)

Application	Model	Valence	Arousal
Application A	8B	6.44 (9.47)	19.93 (15.31)
Application A	70B	6.15 (11.06)	24.29 (13.42)
Application B	8B	5.48 (8.70)	26.35 (14.24)
Application B	70B	8.63 (14.23)	25.09 (18.11)
Overall		6.67 (10.78)	23.92 (15.07)

Note. Session means of AFFDEX Adaptive Valence (−100 to 100) and Adaptive Engagement (0 to 100) over all frames with a detected face; $n = 12$ sessions per cell.

Neither valence nor arousal differed significantly between the model sizes or applications, and there were no interactions (all Holm-corrected $p \geq .272$; Table 7.6). The only tendency concerned the sequence: participants who started with Application B showed somewhat lower valence ($\chi^2(1) = 3.33$, $p = .068$) and arousal ($\chi^2(1) = 2.79$, $p = .095$). The residuals of the arousal model were not normally distributed ($W = 0.92$, $p = .003$).

Table 7.6: Likelihood-ratio tests for displayed valence and arousal ($\chi^2(1)$ and Holm-corrected p)

	Model size		Application		Interaction		Sequence	
Outcome	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}
Valence	1.68	.585	0.62	.862	0.56	.862	3.33	.272
Arousal	0.08	1.000	3.11	.312	0.10	1.000	2.79	.312

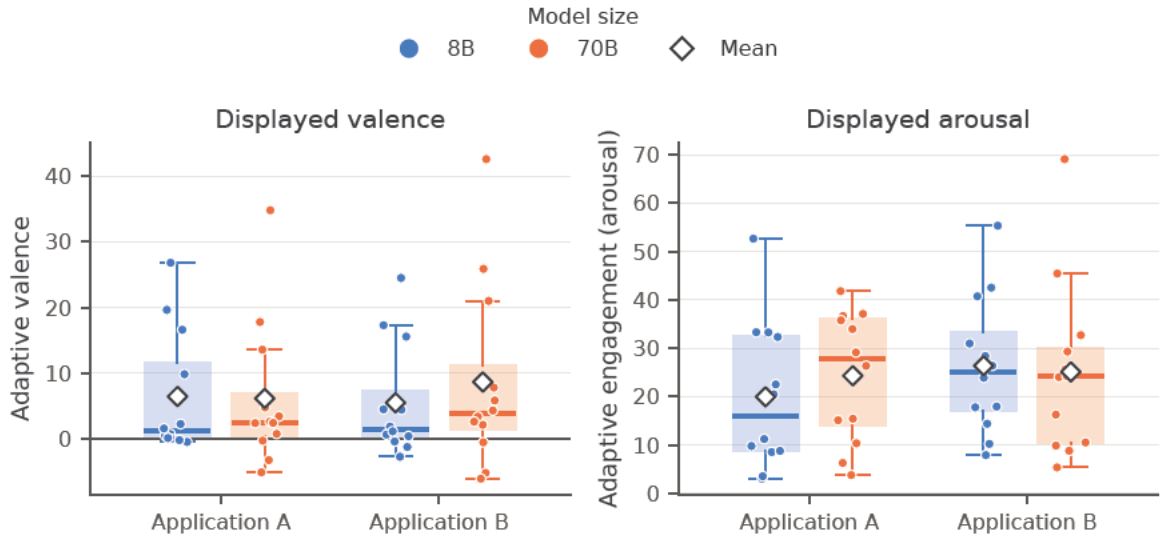

Figure 7.3: Displayed valence and arousal by application and model size (session means of AFFDEX Adaptive Valence and Adaptive Engagement).

7.2.3 UTAUT

Internal consistency was high for the SUS (Cronbach's $\alpha = .82$ for Application A and $\alpha = .94$ for Application B) and for performance expectancy, effort expectancy, social influence and behavioral intention ($\alpha = .90$ to .99; Table 7.7). Facilitating conditions reached only $\alpha = .59$ and .65, so results for this construct should be interpreted with caution.

Table 7.7: Internal consistency (Cronbach's α)

Scale	Items	Application A	Application B
System Usability Scale	10	.82	.94
Performance expectancy (PE)	4	.90	.92
Effort expectancy (EE)	4	.91	.97
Social influence (SI)	2	.97	.99
Facilitating conditions (FC)	4	.59	.65
Behavioral intention (BI)	3	.97	.99

Table 7.8 and Figure 7.4 show the UTAUT constructs on the 7-point scale. Effort expectancy received the highest ratings ($M = 5.88$) and social influence the lowest ($M = 4.17$). Application A was rated higher than Application B on all five constructs, and Application B with the medium model received the lowest rating of the four conditions on every construct.

The residuals of the UTAUT models were normally distributed for PE, SI, FC and BI ($W \geq 0.96$, $p \geq .134$) but not for EE ($W = 0.93$, $p = .006$); the variances were homogeneous for

Table 7.8: UTAUT scores by application × language model, *M* (*SD*)

Application	Model	PE	EE	SI	FC	BI
Application A	8B	4.69 (1.81)	6.12 (1.08)	4.25 (1.92)	5.69 (1.02)	4.58 (2.15)
Application A	70B	5.42 (0.96)	6.35 (0.75)	4.38 (1.84)	5.42 (1.00)	5.50 (1.31)
Application B	8B	5.40 (0.94)	5.92 (1.19)	4.67 (1.98)	5.40 (1.35)	5.50 (1.78)
Application B	70B	4.27 (2.08)	5.10 (2.23)	3.38 (1.91)	5.00 (1.35)	4.08 (2.31)
Overall		4.94 (1.56)	5.88 (1.46)	4.17 (1.91)	5.38 (1.18)	4.92 (1.96)

Note. 7-point scale; $n = 12$ sessions per cell. PE = performance expectancy, EE = effort expectancy, SI = social influence, FC = facilitating conditions, BI = behavioral intention.

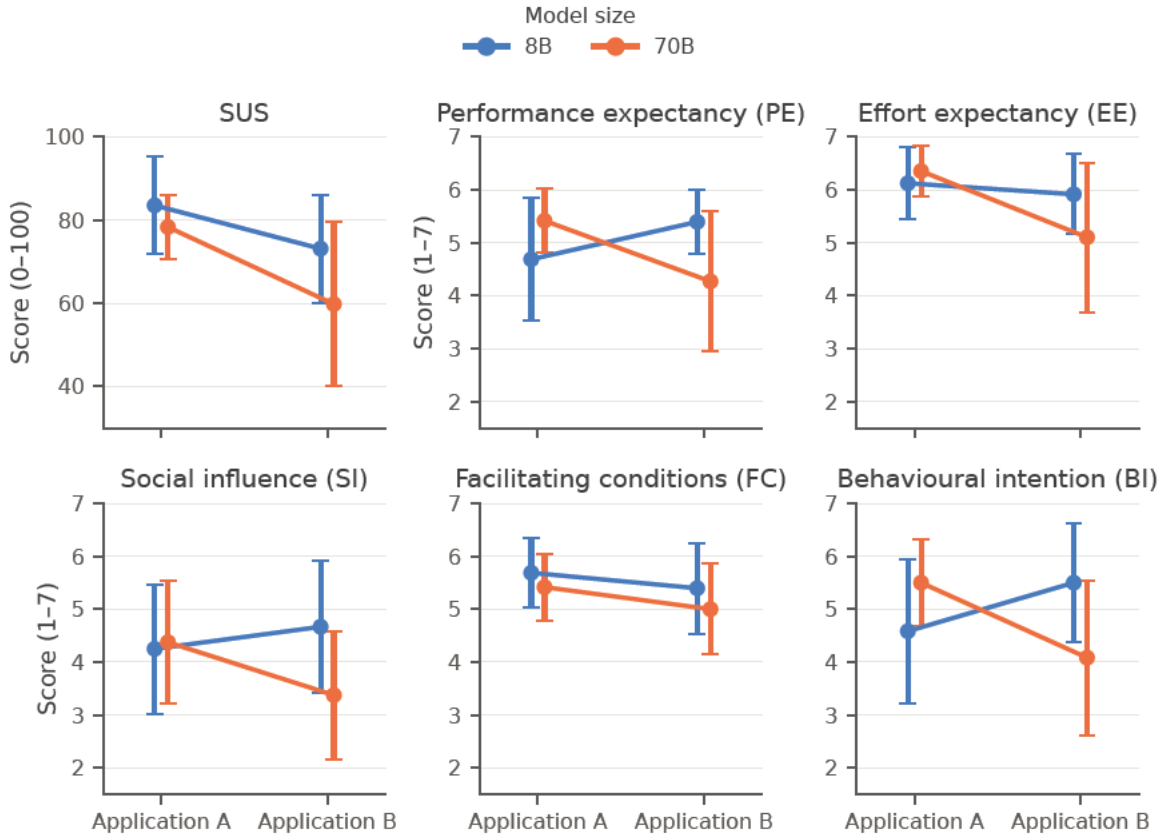

Figure 7.4: Mean SUS and UTAUT scores by application and model size. Error bars show 95 % confidence intervals ($n = 12$ per point).

all constructs ($p \geq .125$). The model size had no significant main effect on any construct (Table 7.9). The small model was rated slightly higher than the medium model on performance expectancy (by 0.23 points, 95 % CI [−0.39, 0.86], $d = 0.15$, $\chi^2(1) = 0.61$, $p = .436$), effort expectancy (by 0.38 points, 95 % CI [−0.29, 1.04], $d = 0.26$, $\chi^2(1) = 1.34$, $p = .246$) and social influence (by 0.19 points, 95 % CI [−0.37, 0.75], $d = 0.10$, $\chi^2(1) = 0.48$, $p = .487$); facilitating conditions and behavioral intention showed the same picture ($\chi^2(1) \leq 1.59$, $p \geq .208$). All Holm-corrected p values were $\geq .623$.

The application had a significant effect on effort expectancy: Application B was rated 0.73 points lower than Application A (95 % CI [−1.26, −0.20], $d = -0.50$, $\chi^2(1) = 7.05$, $p = .008$, Holm-corrected $p = .032$). Facilitating conditions were also rated lower for Application B (by 0.35 points, 95 % CI [−0.66, −0.05], $\chi^2(1) = 5.24$, $p = .022$), but this difference was not significant after Holm correction ($p = .088$). The application had no significant effect on PE, SI, or BI ($\chi^2(1) \leq 2.11$, $p \geq .146$).

None of the interactions between model size and application were significant. The largest occurred for performance expectancy ($\chi^2(1) = 2.90$, $p = .089$, Holm-corrected $p = .354$): in Application B the small model was rated higher than the medium model (5.40 vs. 4.27), whereas in Application A the order was reversed (4.69 vs. 5.42; Figure 7.4). For all other constructs the interaction was negligible ($\chi^2(1) \leq 0.69$, $p \geq .406$), and the sequence had no significant effect ($\chi^2(1) \leq 0.33$, $p \geq .569$).

Because one participant (TP2) rated Application B at or near the scale minimum on the SUS (2.5) and on every UTAUT construct except facilitating conditions, the models were re-estimated without this participant. The results for the model size did not change ($p \geq .277$), and Application A still received significantly higher SUS scores ($\chi^2(1) = 8.93$, Holm-corrected $p = .011$). The effect of the application on effort expectancy, however, decreased to 0.69 points and was no longer significant after Holm correction ($\chi^2(1) = 6.02$, $p = .014$, Holm-corrected $p = .057$). Excluding TP7, the influential case in the regression analyses (Section 7.3), did not weaken the effect on EE ($\chi^2(1) = 8.64$, Holm-corrected $p = .013$). Together with the non-normal residuals for EE, this effect should therefore be interpreted with caution.

Table 7.9: Likelihood-ratio tests for the UTAUT constructs ($\chi^2(1)$ and Holm-corrected p)

	Model size		Application		Interaction		Sequence	
Outcome	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}	χ^2	p_{Holm}
PE	0.61	1.000	0.89	1.000	2.90	.354	0.18	1.000
EE	1.34	.739	7.05	.032	0.61	.872	0.14	.872
SI	0.48	1.000	2.11	.585	0.57	1.000	0.33	1.000
FC	1.59	.623	5.24	.088	0.04	1.000	0.18	1.000
BI	0.40	1.000	0.67	1.000	0.69	1.000	0.01	1.000

7.3 Testing Hypotheses

This section tests the hypotheses from Section 4.1, first the relationship between perceived usability and the UTAUT constructs (H1a–H1c) and then the moderated relationship between displayed valence, arousal, and the UTAUT constructs (H2a–H2c). All models were estimated separately for each application ($n = 24$). Additionally, the effects of the application and model size (H3a–H4c) are evaluated.

7.3.1 Usability × Technology Acceptance (H1a, H1b, H1c)

The perceived usability was strongly and positively related to both performance expectancy and effort expectancy in both applications (Table 7.10, Figure 7.5). For performance expectancy (H1a), the standardized slope was $\beta = .65$ in Application A ($b = 0.061$, 95 % CI [0.030, 0.093], $t(22) = 4.01$, $p < .001$, $R^2 = .42$) and $\beta = .77$ in Application B ($b = 0.049$, 95 % CI [0.031, 0.066], $t(22) = 5.75$, $p < .001$, $R^2 = .60$). For effort expectancy (H1b), the slope was $\beta = .76$ in Application A ($b = 0.045$, 95 % CI [0.028, 0.062], $t(22) = 5.47$, $p < .001$, $R^2 = .58$) and $\beta = .92$ in Application B ($b = 0.062$, 95 % CI [0.050, 0.074], $t(22) = 10.88$, $p < .001$, $R^2 = .84$). A difference of 10 SUS points thus corresponded to about half a point on the 7-point UTAUT scale.

The residuals of the H1a models deviated from normality (Shapiro–Wilk $W = 0.80$ and 0.71, both $p < .001$). This was caused by a single participant (TP7) who rated the usability as good (SUS 75.0 and 82.5) but performance expectancy, social influence, and behavioral intention at the scale minimum in both applications. Without this participant, the relationship became even stronger ($\beta = .82$ and .93, $R^2 = .67$ and .86). All four relationships also remained significant without the other influential cases ($p \leq .001$) and in the pooled mixed models ($\beta = .69$ for PE and $\beta = .91$ for EE, both $p < .001$). H1a and H1b are therefore supported.

Table 7.10: Regression of performance expectancy (H1a) and effort expectancy (H1b) on SUS

Hyp.	Application	Sample	n	b	95 % CI	β	t	p	R^2
H1a	Application A	All	24	0.061	[0.030, 0.093]	0.65	4.01	< .001	0.42
		Excl. TP7	23	0.063	[0.043, 0.083]	0.82	6.56	< .001	0.67
H1a	Application B	All	24	0.049	[0.031, 0.066]	0.77	5.75	< .001	0.60
		Excl. TP7	23	0.051	[0.042, 0.061]	0.93	11.30	< .001	0.86
H1b	Application A	All	24	0.045	[0.028, 0.062]	0.76	5.47	< .001	0.58
		Excl. TP2, TP24	22	0.034	[0.016, 0.052]	0.66	3.96	< .001	0.44
H1b	Application B	All	24	0.062	[0.050, 0.074]	0.92	10.88	< .001	0.84
		Excl. TP2, TP3, TP12	21	0.059	[0.044, 0.074]	0.89	8.43	< .001	0.79

Note. OLS per application. b = change in the UTAUT score (1–7) per SUS point. Excl. = without cases with Cook's $D > 4/n$.

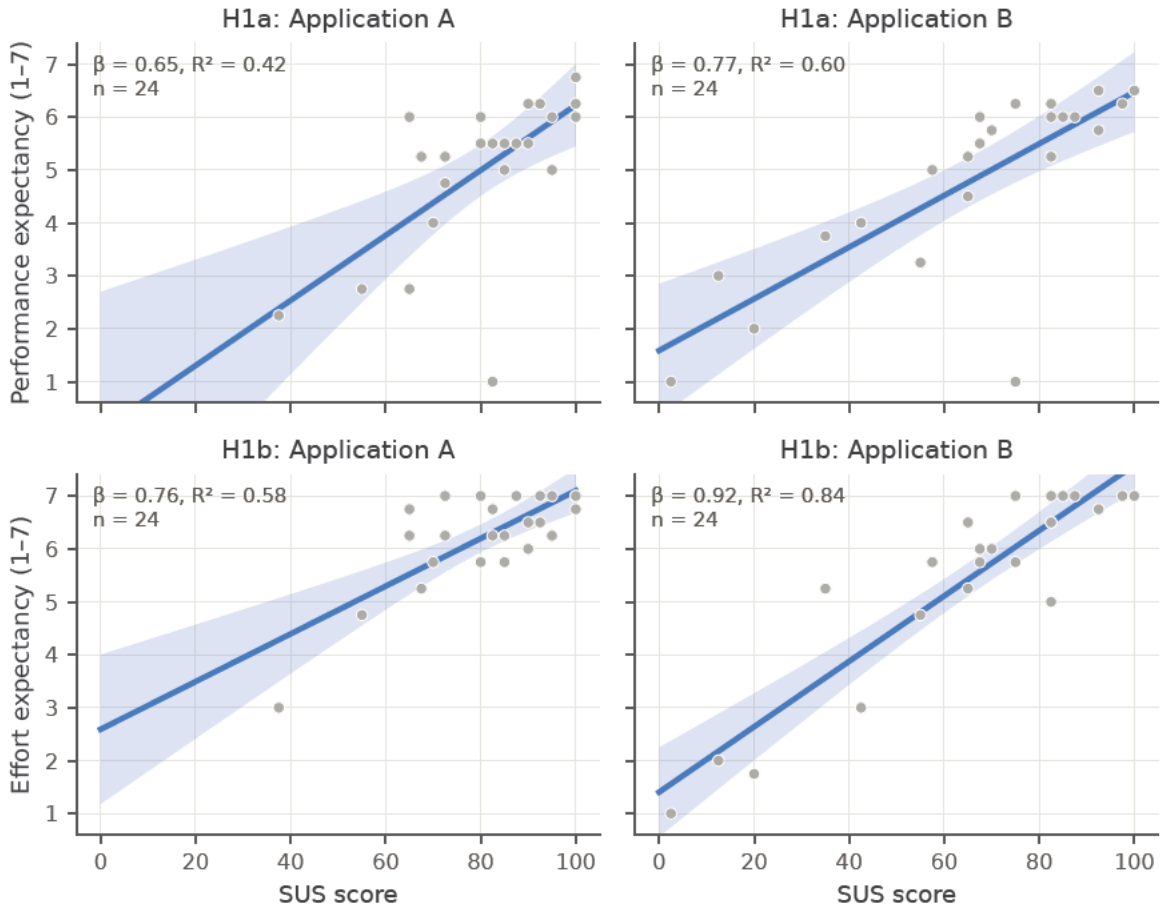

Figure 7.5: Relationship between perceived usability (SUS) and PE (H1a) or EE (H1b), by application. Lines show the Ordinary Least Squares (OLS) fit with its 95 % confidence band.

H1c predicted that perceived usability is more strongly related to effort expectancy than to performance expectancy. Within each application, the correlation of the SUS with effort expectancy ($r_{SUS,EE}$) was therefore compared with its correlation with performance expectancy ($r_{SUS,PE}$), using the test for dependent overlapping correlations by Meng et al. [174] and a confidence interval for the difference $\Delta r = r_{SUS,EE} - r_{SUS,PE}$ after Zou [175] (Table 7.11).

In both applications, the SUS correlated more strongly with effort expectancy than with performance expectancy. In Application A, this difference was small and not significant ($r = .76$ vs. .65, $\Delta r = .11$, 95 % CI [−.14, .41], $z = 0.88$, $p = .378$). In Application B, it was significant ($r = .92$ vs. .77, $\Delta r = .14$, 95 % CI [.01, .37], $z = 2.17$, $p = .030$).

Both differences, however, were produced by TP7, the participant who rated the usability as good but performance expectancy at the scale minimum. This single case lowered $r_{SUS,PE}$ but hardly affected $r_{SUS,EE}$. Without TP7, the difference disappeared in Application B ($r = .92$ vs. .93, $\Delta r = .00$, 95 % CI [−.09, .08], $z = -0.12$, $p = .904$) and reversed its direction in Application A ($r = .76$ vs. .82, $\Delta r = -.06$, 95 % CI [−.29, .13], $z = -0.69$, $p = .490$). Spearman rank correlations, which are less sensitive to single extreme cases, likewise showed no consistent advantage of effort expectancy in the full sample: $\rho_{SUS,EE} = .56$ vs. $\rho_{SUS,PE} = .68$ in Application A and .85 vs. .81 in Application B.

Perceived usability was therefore related to effort expectancy and performance expectancy to a similar extent. The only significant difference depended on a single participant, so H1c is not supported.

Table 7.11: Comparison of the correlations of perceived usability with effort expectancy and performance expectancy (H1c)

Application	Sample	n	$r_{SUS,EE}$	$r_{SUS,PE}$	$r_{EE,PE}$	Δr	95 % CI	z	p
Application A	All	24	0.76	0.65	0.59	0.11	[−0.14, 0.41]	0.88	.378
	Excl. TP7	23	0.76	0.82	0.73	−0.06	[−0.29, 0.13]	−0.69	.490
Application B	All	24	0.92	0.77	0.69	0.14	[0.01, 0.37]	2.17	.030
	Excl. TP7	23	0.92	0.93	0.90	0.00	[−0.09, 0.08]	−0.12	.904

Note. Test for dependent overlapping correlations [174]; 95 % CI after Zou [175]. $\Delta r = r_{SUS,EE} - r_{SUS,PE}$, computed from the unrounded correlations; positive values are in the direction of H1c. TP7 = participant with performance expectancy at the scale minimum despite good SUS scores (influential case in H1a).

7.3.2 Valence and Arousal × Technology Acceptance (H2a, H2b, H2c)

None of the interactions between displayed valence and arousal were significant (Table 7.12, Figure 7.6). In the full samples, adding the interaction increased the explained variance by at most $\Delta R^2 = .10$ (H2b, Application A: $\beta_{V \times A} = -.76$, $t(20) = -1.51$, $p = .147$); in all other full-sample models the increase was $\Delta R^2 \leq .02$ ($p \geq .417$). Without the influential cases,

the largest increase was $\Delta R^2 = .12$ (H2a, Application B, $p = .106$). In the full samples, the interaction coefficients that tended to appear were negative, i.e., opposite to the expected strengthening at higher arousal.

Displayed valence was not positively related to any UTAUT construct either. In Application A, the valence coefficients were close to zero ($\beta = -.20$ to $-.02$, $p \geq .392$). In Application B, they were negative throughout, and significantly so for effort expectancy ($\beta = -.54$, $p = .012$; performance expectancy: $\beta = -.41$, $p = .059$; social influence: $\beta = -.33$, $p = .129$): participants who displayed more positive facial expressions while using Application B rated it as more effortful. This negative relationship, however, was not robust. It disappeared when the influential cases were excluded ($\beta = -.43$, $p = .120$) and was absent in the pooled mixed models (valence $\beta = -.13$ to $-.04$, all $p \geq .619$; interaction $p \geq .148$). Arousal was not significantly related to any construct ($p \geq .090$). Using the raw instead of the adaptive AFFDEX values led to the same pattern, including the negative valence coefficient for effort expectancy in Application B ($\beta = -.50$, $p = .019$). Hypotheses H2a, H2b, and H2c are therefore not supported.

Table 7.12: Hierarchical moderated regression of PE (H2a), EE (H2b), and SI (H2c) on displayed valence and arousal

				Model 1: valence + arousal					Model 2: + interaction		
Hyp.	Application	Sample	n	β_V	p	β_A	p	R^2	$\beta_{V\times A}$	p	ΔR^2
H2a	Application A	All	24	−0.20	.392	0.27	.243	0.07	−0.31	.554	0.016
		Excl. TP12, TP17	22	0.00	.988	0.28	.224	0.08	−0.28	.533	0.020
H2a	Application B	All	24	−0.41	.059	0.35	.100	0.20	−0.04	.889	0.001
		Excl. TP2, TP4, TP12, TP18	20	−0.44	.086	0.43	.090	0.21	0.52	.106	0.122
H2b	Application A	All	24	−0.02	.948	0.01	.970	0.00	−0.76	.147	0.102
		Excl. TP2	23	−0.12	.616	−0.17	.469	0.06	−0.78	.139	0.105
H2b	Application B	All	24	−0.54	.012	0.19	.339	0.27	−0.22	.417	0.024
		Excl. TP2, TP4, TP12, TP18, TP23	19	−0.43	.120	0.16	.533	0.15	0.08	.804	0.004
H2c	Application A	All	24	−0.06	.802	0.04	.877	0.00	−0.13	.807	0.003
H2c	Application B	All	24	−0.33	.129	0.30	.167	0.14	0.03	.927	0.000
		Excl. TP12	23	−0.20	.371	0.31	.164	0.11	0.13	.654	0.010

Note. OLS per application; valence and arousal grand-mean centered within the application. β = standardized coefficient; ΔR^2 = increase from Model 1 to Model 2. Excl. = without cases with Cook's $D > 4/n$ in Model 2.

7.3.3 Categorical Emotions × Technology Acceptance

To examine whether a categorical view of affect explains acceptance better than valence and arousal, the seven basic emotions recognized by AFFDEX were related to PE, EE, SI, and BI. The displayed emotions were dominated by neutral expressions (Table 7.13). Joy reached the highest evidence on average ($M = 6.73$, $Mdn = 3.01$ on a scale from 0 to 100) and contempt the lowest ($M = 0.32$). Fear and surprise were close to zero in most sessions but reached values of up to 36 and 34 in single sessions.

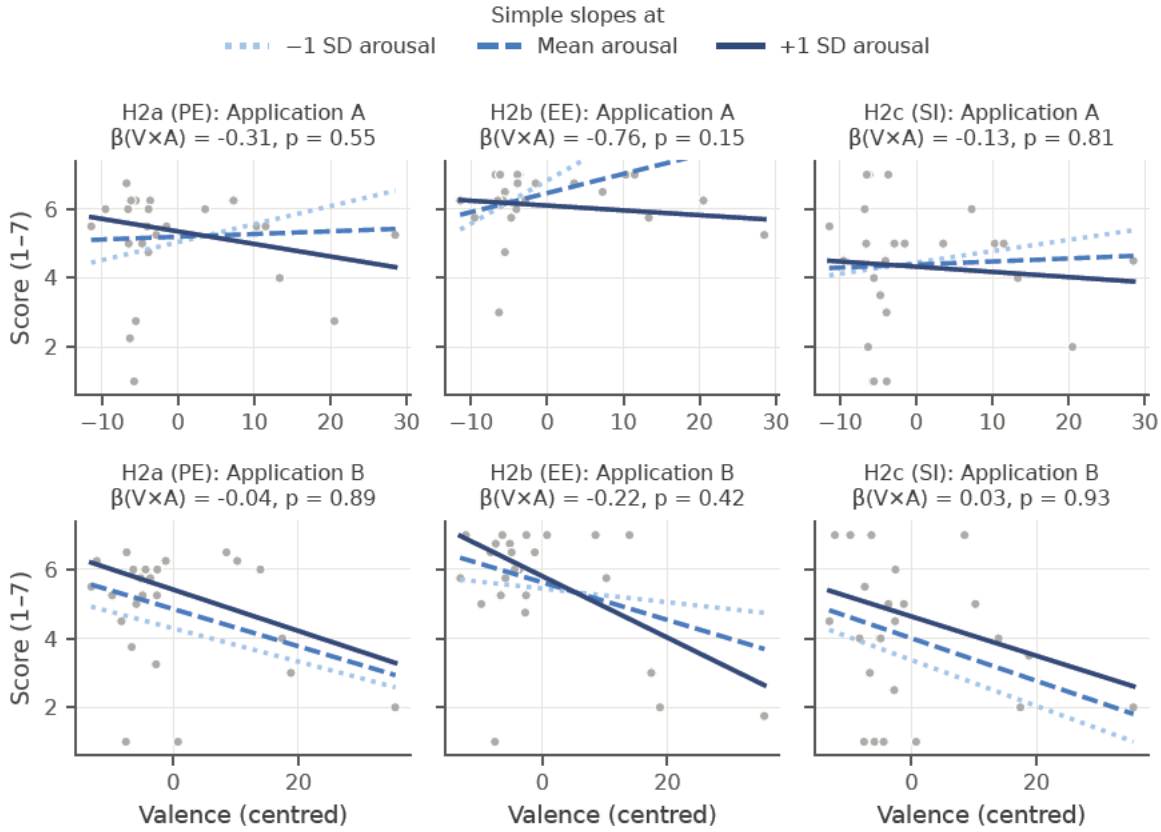

Figure 7.6: Moderated regression of performance expectancy (H2a), effort expectancy (H2b), and social influence (H2c) on displayed valence by application. Lines show the predicted relationship at low (−1 *SD*), mean, and high (+1 *SD*) arousal; dots are the individual sessions.

Table 7.13: Displayed basic emotions by application, *M* (*SD*) and *Mdn*

	Application A		Application B	
Emotion	*M* (*SD*)	*Mdn*	*M* (*SD*)	*Mdn*
Anger	1.09 (1.56)	0.38	1.39 (1.90)	0.56
Contempt	0.31 (0.21)	0.27	0.32 (0.20)	0.32
Disgust	0.64 (0.74)	0.35	1.13 (1.16)	0.71
Fear	2.76 (5.64)	0.60	3.76 (7.97)	0.86
Joy	6.14 (7.07)	2.82	7.31 (8.74)	3.35
Sadness	1.13 (1.34)	0.52	1.09 (1.00)	0.65
Surprise	2.60 (5.29)	0.66	3.77 (7.78)	0.94

Note. Session means of the AFFDEX emotion evidence (0 to 100) over all frames with a detected face; $n = 24$ sessions per application.

None of the basic emotions was significantly related to a UTAUT construct after correction (Table 7.14). In Application A, all correlations were small ($|\rho| \leq .32$, $p \geq .124$). In Application B, several negative emotions correlated *positively* with the ratings: disgust and anger with social influence ($\rho = .52$, $p = .009$ and $\rho = .47$, $p = .019$), and fear and surprise with performance expectancy ($\rho = .41$, $p = .049$ and $\rho = .42$, $p = .044$). None of these coefficients remained significant after Benjamini–Hochberg correction ($q \geq .251$), and their direction is opposite to what would be expected if these expressions reflected a negative experience. Together with the negative valence coefficients in Application B, they rather indicate that participants who were more facially expressive overall rated Application B somewhat higher. Joy, the emotion most closely related to positive valence, was unrelated to all constructs ($|\rho| \leq .21$).

Table 7.14: Spearman correlations between the displayed basic emotions and the UTAUT constructs

	Application A				Application B			
Emotion	PE	EE	SI	BI	PE	EE	SI	BI
Anger	0.26	0.00	0.13	0.29	0.33	0.37	0.47*	0.38
Contempt	0.04	−0.32	−0.05	0.02	0.06	−0.09	0.31	−0.03
Disgust	0.11	0.10	0.07	0.16	0.35	0.27	0.52*	0.35
Fear	0.30	−0.06	−0.09	0.27	0.41*	0.26	0.21	0.38
Joy	−0.02	0.04	0.00	−0.06	0.07	0.02	−0.04	−0.21
Sadness	0.28	−0.03	0.21	0.31	0.28	0.24	0.39	0.25
Surprise	0.18	−0.14	−0.20	0.15	0.42*	0.21	0.16	0.35

Note. $n = 24$ per application. * $p < .05$ uncorrected; no coefficient remains significant after Benjamini–Hochberg correction within the application ($q \geq .251$).

The mixed models over all 48 sessions confirmed this picture. No single emotion predicted a construct after Holm correction (all $p_{\text{Holm}} \geq .295$; Table A.9 in the appendix); the strongest relationship was a positive one between sadness and behavioral intention ($\chi^2(1) = 4.13$, $p = .042$). Entered together, the seven emotions did not improve the prediction of any construct beyond the design factors ($\chi^2(7) \leq 10.61$, $p \geq .156$; Table 7.15). The marginal R^2 increased by .06 to .21, which is about what seven unrelated predictors would add with 48 sessions, and the leave-one-out R^2 was negative for every construct in both applications (−1.02 to −0.44), i.e., predictions from the emotions were less accurate than the mean. Thus, the categorical emotions neither had a significant effect on technology acceptance nor could they predict it.

Table 7.15: Prediction of the UTAUT constructs by all seven basic emotions

	Mixed model (48 sessions)					LOO R^2	
Outcome	$\chi^2(7)$	p	p_{Holm}	R^2_m design	R^2_m + emotions	Application A	Application B
PE	4.63	.705	1.000	.02	.13	−0.77	−0.72
EE	10.61	.156	.626	.09	.30	−1.02	−0.48
SI	4.38	.735	1.000	.02	.08	−0.69	−0.49
BI	10.20	.177	.626	.01	.22	−0.44	−0.77

Note. Likelihood-ratio test of the seven log-transformed, standardized emotions against the model with application, model size, and sequence only (random intercept per participant). R^2_m = marginal R^2 [178]. LOO R^2 = leave-one-out R^2 of an OLS regression on the seven emotions per application ($n = 24$); negative values indicate predictions worse than the mean.

7.3.4 Application and Model Size (H3a–H4c)

Table 7.16 summarizes the main effects of the application and the model size from the mixed models in Section 7.2.

Application A received significantly higher SUS scores than Application B (by 14.5 points, $p_{\text{Holm}} = .006$), and this difference remained significant without the participant with extreme ratings (TP2, $p_{\text{Holm}} = .011$). H3a is therefore supported. Neither displayed valence nor arousal differed between the applications ($p_{\text{Holm}} \geq .312$), so H3b is not supported. Of the five UTAUT constructs, only effort expectancy differed significantly between the applications, being 0.73 points lower for Application B ($p_{\text{Holm}} = .032$). Facilitating conditions showed the same direction but were not significant after correction ($p_{\text{Holm}} = .088$), and PE, SI, and BI did not differ ($p_{\text{Holm}} \geq .585$). As the difference in effort expectancy was no longer significant without TP2 ($p_{\text{Holm}} = .057$), H3c is only partially supported, and this partial support is not robust.

The model size had no significant effect on any measure. The SUS score was 7.9 points higher for the small model, but not significantly so ($p_{\text{Holm}} = .375$), and the differences in displayed valence and arousal ($p_{\text{Holm}} \geq .585$) and in the UTAUT constructs ($p_{\text{Holm}} \geq .623$) were small and not significant. H4a, H4b, and H4c are therefore not supported. With $n = 12$ sessions per cell, however, only large differences could be detected, so these results do not show that the two model sizes are equivalent.

7.4 Exploratory Findings

To explore whether the ratings were related to how the participants used the applications or to their characteristics, Spearman rank correlations were computed between the usage

Table 7.16: Tests of the differences between the applications (H3) and the model sizes (H4)

Hyp.	Measure	Difference	95 % CI	$\chi^2(1)$	p_{Holm}	Result
H3a	SUS	−14.48	[−22.96, −6.00]	10.04	.006	Supported
H3b	Valence	0.77	[−1.25, 2.79]	0.62	.862	Not supported
	Arousal	3.61	[−0.52, 7.74]	3.11	.312	
H3c	PE	−0.22	[−0.70, 0.26]	0.89	1.000	Partially supported
	EE	−0.73	[−1.26, −0.20]	7.05	.032	(EE only,
	SI	−0.29	[−0.70, 0.12]	2.11	.585	not robust)
	FC	−0.35	[−0.66, −0.05]	5.24	.088	
	BI	−0.25	[−0.88, 0.38]	0.67	1.000	
H4a	SUS	7.88	[−2.65, 18.42]	2.35	.375	Not supported
H4b	Valence	−1.75	[−4.52, 1.02]	1.68	.585	Not supported
	Arousal	−0.77	[−6.24, 4.70]	0.08	1.000	
H4c	PE	0.23	[−0.39, 0.86]	0.61	1.000	Not supported
	EE	0.38	[−0.29, 1.04]	1.34	.739	
	SI	0.19	[−0.37, 0.75]	0.48	1.000	
	FC	0.25	[−0.16, 0.66]	1.59	.623	
	BI	0.25	[−0.57, 1.07]	0.40	1.000	

Note. Main effects from the linear mixed-effects models (48 sessions; Tables 7.4, 7.6, and 7.9). Difference = Application B − Application A (H3) and 8B − 70B (H4) in the units of the measure, with Wald 95 % CI from the REML fit. p_{Holm} = likelihood-ratio test, Holm-corrected across the four terms per outcome.

metrics, the demographic items, and the SUS and UTAUT scores, separately for each application ($n = 24$). Because 120 coefficients were inspected and no hypotheses had been formulated, the results are reported descriptively. The p values are uncorrected, and none of the coefficients remained significant after Benjamini–Hochberg correction [176]. With $n = 24$, only correlations of about $|\rho| \geq .40$ reach $p < .05$; about six of the 120 coefficients would be expected to do so by chance alone. The patterns below should therefore be treated as starting points for future studies rather than as findings.

7.4.1 Demographics × Technology Acceptance

The clearest pattern concerned prior knowledge of tenancy law (Figure 7.7, Table 7.17). Participants who knew more about tenancy law rated both applications less favorably on almost every measure ($\rho = -.42$ to $-.03$), most clearly for social influence ($\rho = -.42$, $p = .041$) and effort expectancy ($\rho = -.38$, $p = .067$) in Application A. Affinity for digital applications was positively related to behavioral intention in both applications ($\rho = .28$ and .30) and to effort expectancy in Application B ($\rho = .35$, $p = .093$). Participants who used chatbots more frequently reported less social influence ($\rho = -.35$ and $-.34$). Gender, age, education, and German proficiency showed only small and inconsistent associations.

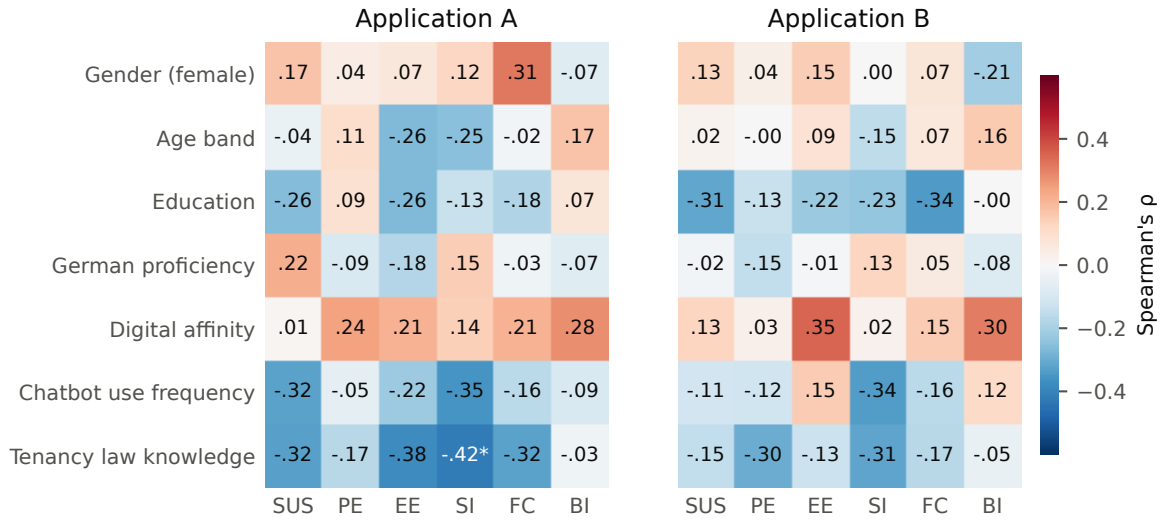

Figure 7.7: Spearman correlations between demographic characteristics and the SUS and UTAUT ratings, by application ($n = 24$; education $n = 23$). German proficiency and chatbot use were reverse-coded so that higher values indicate better proficiency and more frequent use; gender coded female = 1. * $p < .05$, uncorrected.

Table 7.17: Spearman correlations between demographics and the SUS / UTAUT ratings

	Application A						Application B					
	SUS	PE	EE	SI	FC	BI	SUS	PE	EE	SI	FC	BI
Gender (female)	0.17	0.04	0.07	0.12	0.31	−0.07	0.13	0.04	0.15	0.00	0.07	−0.21
Age band	−0.04	0.11	−0.26	−0.25	−0.02	0.17	0.02	0.00	0.09	−0.15	0.07	0.16
Education	−0.26	0.09	−0.26	−0.13	−0.18	0.07	−0.31	−0.13	−0.22	−0.23	−0.34	0.00
German proficiency	0.22	−0.09	−0.18	0.15	−0.03	−0.07	−0.02	−0.15	−0.01	0.13	0.05	−0.08
Digital affinity	0.01	0.24	0.21	0.14	0.21	0.28	0.13	0.03	0.35	0.02	0.15	0.30
Chatbot use frequency	−0.32	−0.05	−0.22	−0.35	−0.16	−0.09	−0.11	−0.12	0.15	−0.34	−0.16	0.12
Tenancy law knowledge	−0.32	−0.17	−0.38	−0.42*	−0.32	−0.03	−0.15	−0.30	−0.13	−0.31	−0.17	−0.05

Note. $n = 24$ per application (education: 23). German proficiency and chatbot use were reverse-coded so that higher values indicate better proficiency and more frequent use; gender coded female = 1. * $p < .05$ uncorrected; no coefficient remains significant after Benjamini–Hochberg correction [176].

7.4.2 Application Usage × Technology Acceptance

The usage metrics were related to the ratings in opposite directions in the two applications (Figure 7.8, Table 7.18). In Application B, longer sessions, more LLM responses, and longer responses all went along with lower ratings: participants who spent more time in Application B rated its usability ($\rho = -.34$) and performance expectancy ($\rho = -.38$, $p = .070$) lower, and the number of responses correlated with the SUS score at $\rho = -.33$ ($p = .118$). Longer responses in Application B went along with lower ratings on every measure ($\rho = -.29$ to $-.20$). In Application A, these associations were close to zero or slightly positive (response length: $\rho = .00$ to $.29$).

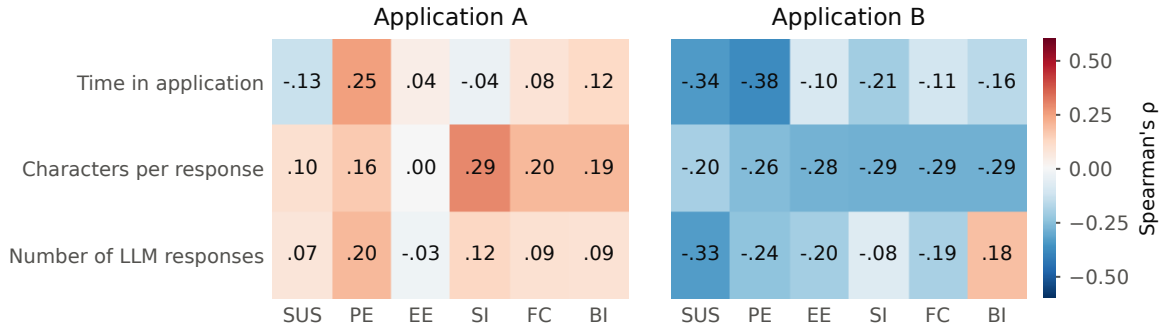

Figure 7.8: Spearman correlations between usage metrics and the SUS and UTAUT ratings, by application ($n = 24$; characters per response: $n = 23$ and 19)

Table 7.18: Spearman correlations between usage metrics and the SUS / UTAUT ratings

	Application A						Application B					
	SUS	PE	EE	SI	FC	BI	SUS	PE	EE	SI	FC	BI
Time in application	−0.13	0.25	0.04	−0.04	0.08	0.12	−0.34	−0.38	−0.10	−0.21	−0.11	−0.16
Characters per response	0.10	0.16	0.00	0.29	0.20	0.19	−0.20	−0.26	−0.28	−0.29	−0.29	−0.29
Number of LLM responses	0.07	0.20	−0.03	0.12	0.09	0.09	−0.33	−0.24	−0.20	−0.08	−0.19	0.18

Note. $n = 23$ per application; characters per response only for sessions with at least one LLM reply (Application A: 23, Application B: 19).

8 Discussion

This chapter interprets the results in light of the hypotheses and the related work (Section 8.1) and answers the research questions (Section 8.2). It then derives refinements of the UVA-UTAUT (Section 8.3), discusses the implications for research and for the design of AIVAs (Section 8.4), and closes with the limitations of the study (Section 8.5).

8.1 Evaluating Hypotheses

The following paragraphs evaluate the hypotheses on perceived usability (H1) and displayed affect (H2) before discussing the effects of the application (H3) and model size (H4) and the role of answer quality. Table 8.1 gives an overview of all results.

Perceived Usability Perceived usability, measured through the SUS, was strongly related to both performance expectancy and effort expectancy in both applications, supporting H1a and H1b. This is in line with Lah et al. [18], who found that substituting the SUS for perceived ease of use in a TAM configuration yielded almost identical explained variance, and with the UX-TAM [127], where perspicuity and dependability predicted PEOU, which in turn predicted perceived usefulness. The results also replicate the pilot study of the UVA-UTAUT [25], in which the standardized coefficients were $\beta = .71$ for PE and $\beta = .76$ for EE with six participants. In this study, they ranged from .65 to .77 for PE and from .76 to .92 for EE, and they were statistically significant in all four models. The usability pathways therefore held although the sample, the application, and the UTAUT item set (17 instead of 8 items) differed, and although usability was rated higher than in the pilot ($M = 73.7$ vs. 53.3). However, the strength of these relationships is likely inflated to some extent by common-method variance, as the SUS and the UTAUT items were completed directly after each other in the same questionnaire.

Regarding H1c, perceived usability was not more strongly related to effort expectancy than to performance expectancy. The only significant difference, in Application B, depended on a single participant (TP7), and without this participant, the difference disappeared in Application B and reversed its direction in Application A. This contrasts with Lah et al. [18], in

whose data the SUS correlated more strongly with perceived ease of use than with perceived usefulness. One possible explanation is that users judge the quality of the answers at the same time as the usability of an AIVA, whereas in conventional software, the interface and the result of the interaction can be separated more clearly. Users may therefore perceive an AIVA as one coherent service rather than as an interface with a function. A second, methodological explanation is a ceiling effect, as in Application A, half of the sessions rated effort expectancy at 6.5 or higher on the 7-point scale ($M = 6.24$, $SD = 0.92$), whereas performance expectancy varied more ($M = 5.05$, $SD = 1.47$). The restricted range of effort expectancy limits how strongly it can correlate with the SUS, so a small difference in the direction of H1c cannot be ruled out with the present sample.

Displayed Affect Displayed valence was not positively related to any UTAUT constructs, and arousal did not strengthen this relationship. Therefore, H2a, H2b, and H2c are not supported.

This also applies to the most promising affective pathway of the pilot study, the valence × arousal interaction on PE ($\Delta R^2 = .21$, $f^2 = 0.90$) [25]. Although the sample of this study was larger than the estimated minimum of 17 participants, the interaction explained almost no additional variance ($\Delta R^2 = .016$ in Application A and .001 in Application B), and its sign was negative. As effect sizes estimated from six participants are highly unstable, the pilot result was most likely overestimated.

Earlier studies that found effects of affect on acceptance relied on self-reported affect measures. In the CAT model, self-reported pleasure and arousal predicted attitude toward a technology [126]. Facial expressions can be more weakly related to experienced emotions, with Barrett et al. [64] finding that the predicted facial expression was observed in only 22 % of reported emotion episodes. Additionally, automated facial expression analysis is considerably less accurate for natural and subtle expressions than for posed ones, with an earlier AFFDEX version correctly identifying only 17 % of positive facial valence in such conditions [71]. The low and mostly neutral displayed valence in this study ($Mdn = 2.39$ on a scale from −100 to 100) reinforces that a solitary interaction with a text-based assistant offers little occasion for expressive behavior. The laboratory setting might have further inhibited participants' expression of affect.

The assumed moderating effect of higher arousal may also be too simple, as Mehrabian and Russell [58] proposed an inverted-U relationship between arousal and approach behavior, meaning that approach behavior decreases again beyond a moderate level of arousal. Arousal is also not a uniform amplifier of valence, as shown in a study by Thüring and Mahlke [86], in which the system with poor usability led to lower self-reported valence but higher arousal. A well-working system thus may be characterized by high valence and low arousal, and a poorly working system by low valence and high arousal. In that case, higher arousal would

not strengthen a positive relationship between valence and acceptance, which is compatible with the negative interaction tendencies found in this study. In addition, positive facial expressions are not necessarily signs of positive experience, as Thüring and Mahlke [86] found the activity of the zygomaticus muscle (also called the "smile muscle") to be higher for the system with poor usability. Barrett et al. [64] also note that smiles track social situations more closely than happiness. This may explain the negative, though not robust, relationship between displayed valence and effort expectancy in Application B.

The study worked with a fictitious scenario in which each participant assumed the role of a landlord with the goal of conducting a modernization measure. This role-play could also have reduced the display of affect, as participants had no real stakes in it. Participants' behavior and ratings may furthermore have been shaped by demand characteristics, that is, by cues about what is expected of them in an experiment [179]. Facial displays in particular depend on the presence of an audience, even an implicit one [180], so that the presence of the researcher and the camera may have influenced what was displayed. Finally, the UTAUT items were formulated for a workplace context (e.g., "in meiner Arbeit" and "Gehaltserhöhung"), which does not entirely fit the scenario context. This may explain why TP7 rated PE, SI, and BI at the scale minimum in both applications despite good SUS scores, and it may have weakened any relationship between affect and these constructs.

The additional analysis of the seven basic emotions leads to the same conclusion. No emotion was related to a UTAUT construct after correction, the emotions together did not improve the prediction of any construct, and their out-of-sample predictions were worse than the mean. A categorical view of affect therefore does not rescue the affective pathway when affect is measured through facial expressions. This is consistent with the criticism that the basic emotion categories cannot be reliably inferred from facial configurations [64], and with the low base rates in this study, in which every emotion other than joy had a median evidence below 1 on a scale of 0 to 100. The positive correlations between negative emotions such as disgust or anger and the ratings of Application B also argue against interpreting these channels as experienced emotions. They mirror the negative valence coefficients in the same application and suggest that the AFFDEX channels mainly reflected how expressive a participant was in general, rather than how they felt about the application.

Model Size Model size had no significant effect on perceived usability, the displayed affect, or any UTAUT construct, although the medium model wrote substantially longer responses in Application B (70B: 2,512 vs. 8B: 1,389 characters). From a user's perspective, the 8B model was therefore at least not noticeably worse. As the study could only detect large effects, this does not show that the models are equivalent. For a domain-specific assistant whose answers are grounded in retrieved documents, model size may matter less than retrieval quality and prompt design.

Purely based on descriptive results, Application B with the 70B model received the lowest rating on every construct, suggesting that the longer answers did not benefit the users. However, the way the models were run is expected to have affected the results, as the 8B model ran locally, whereas the 70B model was served remotely and took about eight times as long to answer, as measured in the technical validation (*Mdn* 32.4 s vs. 3.9 s; Section 5.6), mainly because of its lower generation speed (17 vs. 94 tokens per second). The study data are consistent with this. In Application B, participants spent a similar amount of time with both models (15.3 vs. 15.9 minutes) but received noticeably fewer responses from the 70B model (1.9 vs. 3.7 per session).

Thus, any advantage of the 70B model in answer quality may have been offset by the longer waiting times, particularly in Application B, where the 70B model wrote the longest answers. Model size and hosting are therefore confounded, and the comparison cannot be attributed to model size alone. Small models also have practical advantages in efficiency and privacy, since they can be run on local infrastructure [120], which is especially important in the data-sensitive context of tenancy law.

Application The application had a larger effect than the model, with participants taking 5.3 minutes longer to finish Application B. Application A also received clearly higher SUS scores (80.9 vs. 66.5), a difference that held in 17 of 24 participants and remained significant when single participants were excluded. Application B was also rated lower on effort expectancy, although this difference was no longer significant after one participant with extreme ratings (TP2) was excluded. That the clear difference in perceived usability did not carry over to performance expectancy and behavioral intention suggests that participants separated how easy an application was to use from how useful they considered it for the task, and that both applications were seen as similarly useful. One interpretation is that a long interaction with the open Basic Chat reflects difficulty in obtaining a useful answer rather than engagement, whereas the Guided Assistant structures the interaction regardless of its length. In Application B, longer responses went along with lower ratings on every measure, suggesting that longer answers are not necessarily better answers. Additionally, participants with more self-reported knowledge of tenancy law tended to rate the AIVAs lower than those without or with little knowledge. Prior knowledge may make it easier to notice inaccurate or incomplete answers, although these correlations are exploratory and no participant rated their knowledge as high. Three aspects limit the comparison of the two applications. First, the fallback command was used in 6 of the 24 sessions of Application A (two with the 8B and four with the 70B model). It is only used when the application gets stuck, so it is possible that these errors have negatively influenced the participants' perception of Application A. Second, 4 of the 24 sessions of Application B contained no LLM response (Table 7.1) as they did not use the chat function at all, and for one further session (TP1) no conversation log exists. These participants rated the application without having received an answer from the assistant. Third, the questionnaires

referred to “die Plattform”, so it is unclear whether participants rated the assistant alone or, especially in Application B, the form application around it. The last two limitations are direct consequences of the study design based on the IntelMOD project, as using the web platform was necessary, and it was not possible to force the participants to use the chat function in Application B.

Answer Quality and Retrieval Answer quality was not assessed in the user study, but the technical validation points to weaknesses that may have influenced the ratings. Retrieval was dominated by the dissertation on functional cost splitting, which provided 76 of the 100 retrieved passages, while the BGH rulings were rarely retrieved. This can be partly explained by the composition of the knowledge base, in which the dissertation accounts for 300 of the 480 pages. Page-sized chunks of up to 500 words may also be poorly suited to legal texts, in which short and densely formulated passages carry the relevant content. In addition, the embedding model only processes the beginning of longer passages, so retrieval depends strongly on the first sentences of each chunk. Although RAG is intended to reduce hallucinations, they remain possible [114]. Only 32 % of the 8B answers and 41 % of the 70B answers referred to a document that had actually been retrieved, and some answers named sources that were not part of the retrieved context. Relying on the model to phrase references is therefore not sufficient. A more reliable approach would be to pass structured metadata such as author, year, and page with each chunk and to render the reference deterministically on the server, as is already done for the cost calculation.

8.2 Answering the Research Questions

Based on the evaluation of the hypotheses, the four research questions from Chapter 1 can be answered as follows.

RQ1: How does the perceived usability of an AIVA relate to its acceptance? Perceived usability was strongly and positively related to both performance expectancy and effort expectancy in both applications (H1a and H1b), with standardized coefficients between .65 and .92. It is therefore a central correlate of the acceptance of AIVAs, and the usability pathways of the UVA-UTAUT were supported. Contrary to H1c, perceived usability was not more closely related to effort expectancy than to performance expectancy. For AIVAs, perceived usability thus appears to reflect not only the ease of the interaction but also the usefulness of the answers. As both questionnaires were completed directly after each other, the strength of these relationships may be inflated by common-method variance.

Table 8.1: Overview of the hypotheses and their results

RQ	Hyp.	Statement	Result
RQ1	H1a	Perceived usability is positively related to PE	Supported
	H1b	Perceived usability is positively related to EE	Supported
	H1c	The relationship with EE is stronger than with PE	Not supported
RQ2	H2a	Valence is positively related to PE, moderated by arousal	Not supported
	H2b	Valence is positively related to EE, moderated by arousal	Not supported
	H2c	Valence is positively related to SI, moderated by arousal	Not supported
RQ3	H3a	Perceived usability differs between the applications	Supported (A > B)
	H3b	Valence and arousal differ between the applications	Not supported
	H3c	The UTAUT constructs differ between the applications	Partially supported (EE only, not robust)
RQ4	H4a	Perceived usability differs between the model sizes	Not supported
	H4b	Valence and arousal differ between the model sizes	Not supported
	H4c	The UTAUT constructs differ between the model sizes	Not supported

RQ2: How can the affect that users display during the interaction with an AIVA contribute to explaining its acceptance? In this study, displayed affect did not contribute to explaining acceptance. Displayed valence was not positively related to performance expectancy, effort expectancy, or social influence, and arousal did not strengthen these relationships (H2a–H2c). The categorical view of affect led to the same result, as the seven basic emotions neither had a significant effect on the UTAUT constructs nor were able to predict them, and the few tendencies that appeared pointed in the opposite direction of what would be expected. This does not mean that affect is irrelevant to acceptance, since self-reported affect has predicted acceptance-related constructs in earlier studies [126]. Rather, the answer to RQ2 is that affect can only contribute to explaining acceptance if it is measured validly. Session-level averages of facial expressions recorded during a solitary, text-based interaction mainly captured how expressive a participant was and not how they experienced the AIVA. A contribution of affect would therefore have to be established with self-reported or physiological measures or with event-related analyses of the reactions to individual answers (Section 8.3).

RQ3: Do perceived usability, displayed affect, and acceptance differ between a guided and an open AIVA conversation design? Perceived usability differed clearly, as the Guided Assistant was rated 14.5 SUS points higher than the Basic Chat and participants spent 5.3 minutes less using it (H3a). Acceptance differed only partially, as the Guided Assistant was rated higher on effort expectancy, but this difference was not robust, and performance expectancy, social influence, facilitating conditions, and behavioral intention did not differ (H3c). The displayed affect did not differ (H3b). Guiding users through a complex procedure thus improves the perceived usability of an AIVA, but in this study, this did not translate into a higher intention to use.

RQ4: Do perceived usability, displayed affect, and acceptance differ between an AIVA based on a small (8B) and on a medium-sized (70B) language model? No significant differences between the model sizes were found for perceived usability, displayed affect, or any UTAUT construct (H4a–H4c). Descriptively, the small model was even rated slightly higher. From the users' perspective, the locally hosted 8B model was therefore not noticeably worse than the 70B model. As the study could only detect large effects and model size was confounded with hosting and response time, this answer is limited to the absence of a large difference and does not establish equivalence.

8.3 Model Refinement

The results suggest several refinements of the UVA-UTAUT. The usability pathways are the best-supported part of the model and should be retained. The assumption that perceived usability is more closely related to EE than to PE (H1c) should be dropped, at least for conversational systems.

The affective part of the model requires a more fundamental revision. Since neither the displayed valence nor the displayed basic emotions were related to acceptance in this study, facial expressions should not be the only affective predictor. Self-reported affect, for example, through the SAM [60], directly captures the subjective feeling [52] and has been shown to predict acceptance-related constructs [126]. Electrodermal activity could complement it as a physiological measure of arousal [86]. The linear moderation by arousal should be either removed or replaced by a model that allows for a curvilinear effect of arousal [58] and for its different meaning at low and high valence.

The pathway from affect to social influence should be removed. It explained no meaningful variance in the pilot [25] or in this study ($\Delta R^2 \leq .003$ in the full samples for the interaction), and it lacks a strong theoretical basis, as SI reflects the perceived opinions of others rather than the experience of use. Social influence itself appears less relevant in this context. It received the lowest ratings of all constructs, and in the original UTAUT studies, the social constructs were only significant in mandatory settings [11], whereas the use of a modernization assistant is voluntary.

Finally, constructs specific to AI-based advice should be considered. Trust has repeatedly been found to be relevant for the acceptance of AI-based assistants [135], [137], and the exploratory finding on tenancy law knowledge suggests that domain knowledge may also shape how an AIVA is evaluated, indicating that it could be used as a moderator. The item wording should be adapted to private and voluntary use and should name the assistant instead of a platform.

8.4 Implications

The findings have implications both for research on technology acceptance and affective computing and for the practical design of AIVAs.

Implications for HCI and Affective Computing For research on technology acceptance, the results confirm that perceived usability correlates strongly with both PE and EE, addressing the open question of how the SUS relates to acceptance constructs [19]. For affective

computing, the results do not support FER as a sole predictor of technology acceptance, as displayed valence was low and was not positively related to acceptance, and the basic emotions could not predict it either. This is consistent with the view that facial movements must not be confused with emotions [64]. Session-level averages of facial expressions may be of limited use in this type of study, and future acceptance research should not treat them as a substitute for self-reported affect.

Implications for AIVA Design Several implications follow for the design of AIVAs in high-stakes domains. First, perceived usability should be treated as a central design goal, since it was the strongest correlate of acceptance in both applications. Second, small, locally hosted models appear to be a viable option from the users' perspective. No disadvantage of the 8B model was detected, which supports data-sovereign deployment in line with the GDPR [24]. Third, guiding the user through a complex procedure may reduce the perceived effort compared to an open chat, although this effect was not robust. Fourth, longer answers should not be equated with better answers, and answer length should therefore be limited, for example, through the prompt. Fifth, reliable grounding and verifiable sources are essential, because lay users may not be able to detect errors and are susceptible to automation bias [115]. For a production setting, further tests are strongly recommended to assess the trade-off of enforcing stricter guardrails (Section 5.4).

8.5 Limitations

The results should be interpreted with regard to several limitations concerning the sample and study design, the measures, the emotion recognition, the system itself, and the data quality. Limitations that concern the comparison of the two applications in particular have already been discussed in Section 8.1.

Sample and Study Design Participants were recruited through a newsletter, a networking event, personal contacts, and the university context, which introduces self-selection. The sample was young, highly educated, and experienced with chatbots, which is relevant because SUS scores are sensitive to users' experience [19]. With $n = 24$ per application, only large effects could be detected reliably (Section 6.1), so the non-significant results must not be interpreted as evidence for the absence of an effect. This applies in particular to the moderation hypotheses and to the comparison of the model sizes. Half of the sequences used the same model in both tasks, so the model comparison was partly between subjects. Model size was confounded with hosting, as the 8B model ran locally and the 70B model was served remotely at a considerably lower generation speed. The fallback command replaced the participants'

inputs with a sample building in 6 of the 24 Application A sessions, unequally distributed between the models (two with the 8B and four with the 70B model). Participants took on a fictitious role in a single session, so the results reflect initial acceptance only. The UTAUT is itself essentially a static theory [45], and the determinants of acceptance differ between inexperienced and experienced users [33].

Measures The software is aimed at individuals and, in some cases, owners' associations, so the UTAUT may not always be applicable, as it was developed for organizational settings [11]. The items used here refer to work and salary, and facilitating conditions reached only low reliability, while social influence consisted of only two items. Both questionnaires referred to "die Plattform", which may have caused confusion about whether the assistant or the surrounding application was being rated. As all acceptance and usability measures were self-reported in the same questionnaire block, common-method variance may have inflated their relationships.

Emotion Recognition Affect was measured through a single modality, as eye tracking could not be used because of prolonged technical problems in the laboratory, and no physiological signals were recorded. Multimodal systems outperform unimodal ones only modestly [65], so the absence of a self-report measure of affect is the more serious limitation. Facial expressions are only weakly related to experienced emotion [64], AFFDEX is less accurate for subtle, natural expressions [71], and the engagement metric used as arousal reflects expressiveness rather than physiological activation [69]. The basic emotion channels had very low base rates, so that single sessions with strong expressions dominated their variance. Session-level aggregates also hide short affective episodes, such as reactions to a particular answer. Moreover, facial analysis systems can be demographically biased [73], which limits the transferability of the findings.

System Answer quality was not assessed in the user study, so hallucinations could not be related to the ratings. Retrieval was dominated by a single document, and source references were unreliable (Section 5.6).

Data Quality As a result of unforeseen technical difficulties with the laboratory setup, the conversation log of TP1 with the LLM was lost. This session was therefore excluded from the analyses of response counts and response length. Excluding the participant entirely did not change the results for perceived usability or effort expectancy.

9 Conclusion

This chapter summarizes the thesis and its main findings (Section 9.1), briefly recaps the most important limitations of the study, and outlines directions for future work (Section 9.2).

9.1 Summary

The energy-efficient modernization of rental housing is held back not only by its costs, but also by complex technical and legal questions and by disputes between landlords and tenants over how costs are allocated. AI virtual assistants could make this knowledge accessible to laypeople, but they can only do so if users accept them. This thesis therefore examined how perceived usability and displayed affect relate to the acceptance of AIVAs in the context of building modernization, and whether acceptance differs between a guided and an open conversation design and between a small and a medium-sized language model.

Building on the UVA-UTAUT, two RAG-based assistants were developed in the IntelMOD project: a Guided Assistant that leads users through the functional cost splitting process and a Basic Chat that answers questions alongside a form. In a counterbalanced laboratory study with 24 participants, both applications were used with either an 8B or a 70B Apertus model, while facial expressions were analyzed with AFFDEX and perceived usability and acceptance were measured with the SUS and the UTAUT questionnaire.

Regarding RQ1, perceived usability was strongly related to performance expectancy and effort expectancy in both applications (H1a and H1b supported), and to a similar extent for both constructs (H1c not supported). Regarding RQ2, displayed affect did not contribute to explaining acceptance. Displayed valence was not positively related to acceptance, and arousal did not strengthen this relationship (H2a, H2b, and H2c not supported); the seven basic emotions could not predict acceptance either. This is attributed mainly to the limited validity of facial expressions as indicators of experienced emotion in a solitary, text-based interaction. Regarding RQ3, the Guided Assistant received higher usability ratings than the Basic Chat (H3a supported), whereas displayed affect did not differ (H3b not supported) and only effort expectancy differed among the UTAUT constructs, which was not robust (H3c partially supported). Regarding RQ4, model size had no significant effect on perceived

usability, displayed affect, or any UTAUT construct (H4a–H4c not supported), although the 70B model wrote considerably longer answers in the Basic Chat.

The results support perceived usability as a central construct of acceptance of AIVAs and caution against using facial expression analysis as a sole affective predictor. They also indicate that small, locally hosted models are a promising option from the users' perspective, although equivalence could not be shown and model size was confounded with hosting.

Limitations The limitations of the study are discussed in detail in Section 8.5. The most important ones are the small and homogeneous sample, which only allowed large effects to be detected, the confounding of model size with hosting and response time, and the measurement of affect through facial expressions alone, without a self-report measure. In addition, the UTAUT items were worded for a workplace context, and all acceptance and usability measures were collected in the same questionnaire, so their relationships may be inflated by common-method variance.

9.2 Future Work

Future studies should test the revised UVA-UTAUT (Section 8.3) with a larger sample and with the actual target groups of tenants, landlords, and owners' associations, ideally in a field setting and over a longer period of time. Comparing a small and a large model with identical hosting and equivalence testing would clarify whether small models are indeed sufficient. This should be combined with an objective assessment of answer quality, for example with Ragas [181].

On the measurement side, the UEQ [103] could be collected in addition to the SUS. Since it separates pragmatic from hedonic qualities, it would show whether UX as a whole adds explanatory power beyond perceived usability. Collecting the UMUX [96] or the UMUX-LITE [97], as well as an adjective rating scale [182], would strengthen the knowledge base on how usability measures relate to acceptance. A chatbot-specific scale such as the BUS-11 [130] could capture conversational qualities that the SUS does not cover.

Affect should be measured through self-report in addition to facial expressions, complemented by physiological arousal [86]. Instead of session-level averages, event-related analyses time-locked to the LLM responses could reveal how users react to individual answers. Finally, retrieval should be improved by using a chunking strategy suited to legal texts and an embedding model that covers the whole passage, and source references should be generated from structured metadata.

Bibliography

[1] Bundesrepublik Deutschland, *Bundes-Klimaschutzgesetz (KSG)*, Zuletzt geändert durch Art. 1 des Gesetzes vom 15.07.2024, 2019. Accessed: Oct. 27, 2026. [Online]. Available: https://www.gesetze-im-internet.de/ksg/

[2] Umweltbundesamt, "Emissionsdaten 2025 & Projektionsdaten 2026: Pressehintergrundinformation", Umweltbundesamt, Dessau-Roßlau, Mar. 2026. [Online]. Available: https://www.umweltbundesamt.de/system/files/document/UBA%20Emissionsdaten%202025_Projektionsdaten%202026_Hintergrundpapier_2026_03_14.pdf

[3] Bundesverband energieeffiziente Gebäudehülle e.V. (BuVEG) and B+L Marktdaten, "Sanierungsquote 2025: Talfahrt für energetische Gebäudesanierung geht weiter", BuVEG / B+L Marktdaten Bonn, 2026. [Online]. Available: https://buveg.de/sanierungsquote/

[4] S. Ebrahimigharehbaghi, Q. K. Qian, F. M. Meijer, and H. J. Visscher, "Unravelling Dutch homeowners' behaviour towards energy efficiency renovations: What drives and hinders their decision-making?", *Energy Policy*, vol. 129, pp. 546–561, Jun. 2019. DOI: 10.1016/j.enpol.2019.02.046

[5] Bundesrepublik Deutschland, *§ 555b Bürgerliches Gesetzbuch (BGB) – Modernisierungsmaßnahmen*, 2026.

[6] *Bürgerliches Gesetzbuch (BGB) § 559 Mieterhöhung nach Modernisierungsmaßnahmen*, 2026. [Online]. Available: https://www.gesetze-im-internet.de/bgb/__559.html

[7] K. David et al., *KI-gestützte Modernisierung an Mietwohnungsbaubeständen (KIMM)*, 2025. DOI: 10.4119/unibi/3015046

[8] Hochschule für Wirtschaft und Umwelt Nürtingen-Geislingen (HfWU), *IntelMOD*, 2026. Accessed: Sep. 7, 2026. [Online]. Available: https://transpa-rent.de/intelmod/

[9] K. David, "Funktionales Kostensplitting zur Ermittlung von Mieterhöhungen nach energetischen Maßnahmen: Eine Handlungsempfehlung auf Basis theoretischer und empirischer Untersuchungen", Ph.D. dissertation, HafenCity Universität Hamburg, 2019.

[10] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, Sep. 1989. DOI: 10.2307/249008

[11] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward A Unified View", *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, Sep. 1, 2003. DOI: 10.2307/30036540

[12] D. Gursoy, O. H. Chi, L. Lu, and R. Nunkoo, "Consumers acceptance of artificially intelligent (AI) device use in service delivery", *International Journal of Information Management*, vol. 49, pp. 157–169, Dec. 2019. DOI: 10.1016/j.ijinfomgt.2019.03.008

[13] K. Sohn and O. Kwon, "Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products", *Telematics and Informatics*, vol. 47, p. 101 324, Apr. 2020. DOI: 10.1016/j.tele.2019.101324

[14] M. Blut et al., "Meta-Analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): Challenging its Validity and Charting a Research Agenda in the Red Ocean", *Journal of the Association for Information Systems*, vol. 23, no. 1, pp. 13–95, 2022. DOI: 10.17705/1jais.00719

[15] V. Venkatesh, "Adoption and use of AI tools: A research agenda grounded in UTAUT", *Annals of Operations Research*, vol. 308, no. 1–2, pp. 641–652, Jan. 2022. DOI: 10.1007/s10479-020-03918-9

[16] A. Beaudry and A. Pinsonneault, "The Other Side of Acceptance: Studying the Direct and Indirect Effects of Emotions on Information Technology Use", *MIS Quarterly*, vol. 34, no. 4, pp. 689–710, Dec. 1, 2010. DOI: 10.2307/25750701

[17] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots", *International Journal of Human-Computer Studies*, vol. 151, p. 102 630, Jul. 2021. DOI: 10.1016/j.ijhcs.2021.102630

[18] U. Lah, J. R. Lewis, and B. Šumak, "Perceived Usability and the Modified Technology Acceptance Model", *International Journal of Human–Computer Interaction*, vol. 36, no. 13, pp. 1216–1230, Aug. 2020. DOI: 10.1080/10447318.2020.1727262

[19] J. R. Lewis, "The System Usability Scale: Past, Present, and Future", *International Journal of Human–Computer Interaction*, vol. 34, no. 7, pp. 577–590, Jul. 2018. DOI: 10.1080/10447318.2018.1455307

[20] K. Hildebrandt, T. Ortmann, and L. Putzar, "Acceptance, Usability, and Emotions - An Extension of the UTAUT Designed for AI Virtual Assistants in the Context of Housing Modernization", in *Proceedings of the 18th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, Corfu Island Greece: ACM, Jun. 2025, pp. 575–582, ISBN: 979-8-4007-1402-3. DOI: 10.1145/3733155.3734906

[21] J. Kaplan et al., "Scaling Laws for Neural Language Models", *arXiv preprint arXiv:2001.08361*, 2020. DOI: 10.48550/ARXIV.2001.08361

[22] S. Luccioni, Y. Jernite, and E. Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?", in *The 2024 ACM Conference on Fairness Accountability and Transparency*, Rio de Janeiro Brazil: ACM, Jun. 3, 2024, pp. 85–99, ISBN: 979-8-4007-0450-5. DOI: 10.1145/3630106.3658542

[23] International Energy Agency, "Energy and AI", International Energy Agency (IEA), Paris, Apr. 2025. [Online]. Available: https://www.iea.org/reports/energy-and-ai

[24] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88, 2016. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[25] K. Hildebrandt, M. Treu, T. Ortmann, and L. Putzar, "An empirical validation of the UVA-UTAUT model for virtual assistants", in *Pervasive Technologies Related to Assistive Environments: 19th International Conference, PETRA 2026, Crete, Greece, July 12–15, 2026, Proceedings, Part II*, F. Makedon, V. Athitsos, M. Li, and S. Nikanfar, Eds., ser. Communications in Computer and Information Science, Cham: Springer, 2027, ISBN: 978-3-032-38042-5. [Online]. Available: https://link.springer.com/book/9783032380418 in press.

[26] M. Fishbein and I. Ajzen, *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley, 1975, ISBN: 978-0-201-02089-2.

[27] B. H. Sheppard, J. Hartwick, and P. R. Warshaw, "The Theory of Reasoned Action: A Meta-Analysis of Past Research with Recommendations for Modifications and Future Research", *Journal of Consumer Research*, vol. 15, no. 3, pp. 325–343, Dec. 1988. DOI: 10.1086/209170

[28] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models", *Management Science*, vol. 35, no. 8, pp. 982–1003, 1989. DOI: 10.1287/mnsc.35.8.982

[29] V. Venkatesh and H. Bala, "Technology Acceptance Model 3 and a Research Agenda on Interventions", *Decision Sciences*, vol. 39, no. 2, pp. 273–315, May 2008. DOI: 10.1111/j.1540-5915.2008.00192.x

[30] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies", *Management Science*, vol. 46, no. 2, pp. 186–204, Feb. 2000. DOI: 10.1287/mnsc.46.2.186.11926

[31] I. Ajzen, "The theory of planned behavior", *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179–211, Dec. 1991. DOI: 10.1016/0749-5978(91)90020-T

[32] K. Mathieson, "Predicting User Intentions: Comparing the Technology Acceptance Model with the Theory of Planned Behavior", *Information Systems Research*, vol. 2, no. 3, pp. 173–191, Sep. 1991. DOI: 10.1287/isre.2.3.173

[33] S. Taylor and P. Todd, "Assessing IT Usage: The Role of Prior Experience", *MIS Quarterly*, vol. 19, no. 4, pp. 561–570, Dec. 1, 1995. DOI: 10.2307/249633

[34] S. Taylor and P. A. Todd, "Understanding information technology usage: A test of competing models", *Information Systems Research*, vol. 6, no. 2, pp. 144–176, 1995. JSTOR: 23011007. Accessed: Jun. 10, 2026. [Online]. Available: http://www.jstor.org/stable/23011007

[35] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory* (Prentice-Hall Series in Social Learning Theory). Englewood Cliffs, NJ: Prentice-Hall, 1986, ISBN: 0-13-815614-X.

[36] D. R. Compeau and C. A. Higgins, "Computer Self-Efficacy: Development of a Measure and Initial Test", *MIS Quarterly*, vol. 19, no. 2, pp. 189–211, Jun. 1995. DOI: 10.2307/249688

[37] H. C. Triandis, *Interpersonal Behavior*. Monterey, CA: Brooks/Cole, 1977.

[38] R. L. Thompson, C. A. Higgins, and J. M. Howell, "Personal Computing: Toward a Conceptual Model of Utilization", *MIS Quarterly*, vol. 15, no. 1, pp. 125–143, Mar. 1991. DOI: 10.2307/249443

[39] E. M. Rogers, *Diffusion of innovations*, Fifth edition. New York London Toronto Sydney: Free Press, 2003, ISBN: 978-0-7432-5823-4.

[40] G. C. Moore and I. Benbasat, "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation", *Information Systems Research*, vol. 2, no. 3, pp. 192–222, Sep. 1991. DOI: 10.1287/isre.2.3.192

[41] G. C. Moore and I. Benbasat, "Integrating Diffusion of Innovations and Theory of Reasoned Action models to predict utilization of information technology by end-users", in *Diffusion and Adoption of Information Technology*, K. Kautz and J. Pries-Heje, Eds., Boston, MA: Springer US, 1996, pp. 132–146, ISBN: 978-1-4757-4977-9. DOI: 10.1007/978-0-387-34982-4_10

[42] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "Extrinsic and Intrinsic Motivation to Use Computers in the Workplace", *Journal of Applied Social Psychology*, vol. 22, no. 14, pp. 1111–1132, Jul. 1992. DOI: 10.1111/j.1559-1816.1992.tb00945.x

[43] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior* (Perspectives in Social Psychology). New York: Plenum Press, 1985, 371 pp., ISBN: 978-1-4899-2273-1.

[44] V. Venkatesh, "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model", *Information Systems Research*, vol. 11, no. 4, pp. 342–365, Dec. 2000. DOI: 10.1287/isre.11.4.342.11872

[45] V. Venkatesh, J. Thong, and X. Xu, "Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead", *Journal of the Association for Information Systems*, vol. 17, no. 5, pp. 328–376, May 2016. DOI: 10.17705/1jais.00428

[46] Y. K. Dwivedi, N. P. Rana, A. Jeyaraj, M. Clement, and M. D. Williams, "Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a Revised Theoretical Model", *Information Systems Frontiers*, vol. 21, no. 3, pp. 719–734, Jun. 2019. DOI: 10.1007/s10796-017-9774-y

[47] R. Bagozzi, "The Legacy of the Technology Acceptance Model and a Proposal for a Paradigm Shift.", *Journal of the Association for Information Systems*, vol. 8, no. 4, pp. 244–254, Apr. 2007. DOI: 10.17705/1jais.00122

[48] R. W. Picard, *Affective Computing*. The MIT Press, Sep. 24, 1997, ISBN: 978-0-262-28158-4. DOI: 10.7551/mitpress/1140.001.0001

[49] E. Hudlicka, "To feel or not to feel: The role of affect in human–computer interaction", *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 1–32, Jul. 2003. DOI: 10.1016/S1071-5819(03)00047-8

[50] S. Brave and C. Nass, "Emotion in human-computer interaction", in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, USA: L. Erlbaum Associates Inc., 2002, pp. 81–96, ISBN: 0-8058-3838-4.

[51] L. Stark and J. Hoey, "The Ethics of Emotion in Artificial Intelligence Systems", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 3, 2021, pp. 782–793, ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445939

[52] K. R. Scherer, "What are emotions? And how can they be measured?", *Social Science Information*, vol. 44, no. 4, pp. 695–729, Dec. 2005. DOI: 10.1177/0539018405058216

[53] P. Ekman, "An argument for basic emotions", *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, May 1992. DOI: 10.1080/02699939208411068

[54] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.", *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971. DOI: 10.1037/h0030377

[55] R. Plutchik, "A general psychoevolutionary theory of emotion", in *Theories of Emotion*, Elsevier, 1980, pp. 3–33, ISBN: 978-0-12-558701-3. DOI: 10.1016/B978-0-12-558701-3.50007-7

[56] R. Plutchik, "The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice", *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001. JSTOR: 27857503. Accessed: Oct. 27, 2026. [Online]. Available: http://www.jstor.org/stable/27857503

[57] H. Schlosberg, "Three dimensions of emotion", *Psychological Review*, vol. 61, no. 2, pp. 81–88, 1954. DOI: 10.1037/h0054570

[58] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA: MIT Press, 1974, ISBN: 978-0-262-13090-5.

[59] J. A. Russell, "A circumplex model of affect.", *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980. DOI: 10.1037/h0077714

[60] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential", *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994. DOI: 10.1016/0005-7916(94)90063-9

[61] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales.", *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, Jun. 1988. DOI: 10.1037/0022-3514.54.6.1063

[62] R. A. Calvo and S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications", *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan. 2010. DOI: 10.1109/T-AFFC.2010.1

[63] L. Shu et al., "A Review of Emotion Recognition Using Physiological Signals", *Sensors*, vol. 18, no. 7, p. 2074, Jun. 28, 2018. DOI: 10.3390/s18072074

[64] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements", *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019. DOI: 10.1177/1529100619832930

[65] S. K. D'Mello and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems", *ACM Computing Surveys*, vol. 47, no. 3, pp. 1–36, Apr. 16, 2015. DOI: 10.1145/2682899

[66] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto, CA: Consulting Psychologists Press, 1978. DOI: 10.1037/t27734-000

[67] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. E. Kaliouby, "AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit", in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, San Jose California USA: ACM, May 7, 2016, pp. 3723–3726, ISBN: 978-1-4503-4082-3. DOI: 10.1145/2851581.2890247

[68] iMotions A/S, *iMotions – human behavior research software*, version 11, 2026. Accessed: Nov. 5, 2026. [Online]. Available: https://imotions.com/de/

[69] iMotions A/S, *Facial Expression Analysis - Emotion Detection Software*, 2022. Accessed: Nov. 9, 2026. [Online]. Available: https://imotions.com/products/imotions-lab/modules/fea-facial-expression-analysis/

[70] M. Bishay, K. Preston, M. Strafuss, G. Page, J. Turcot, and M. Mavadati, "AFFDEX 2.0: A Real-Time Facial Expression Analysis Toolkit", *arXiv preprint arXiv:2202.12059*, 2022. DOI: 10.48550/ARXIV.2202.12059

[71] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with AFFDEX and FACET: A validation study", *Behavior Research Methods*, vol. 50, no. 4, pp. 1446–1460, Aug. 2018. DOI: 10.3758/s13428-017-0996-1

[72] D. Dupré, E. G. Krumhuber, D. Küster, and G. J. McKeown, "A performance comparison of eight commercially available automatic classifiers for facial affect recognition", *PLOS ONE*, vol. 15, no. 4, S. D'Mello, Ed., e0231968, Apr. 24, 2020. DOI: 10.1371/journal.pone.0231968

[73] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification", in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, PMLR, Feb. 23–24, 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[74] N. Bevan, J. Kirakowski, and J. Maissel, "What is Usability?", in *Human Aspects in Computing: Design and Use of Interactive Systems and Work with Terminals*, H.-J. Bullinger, Ed., Amsterdam: Elsevier, 1991, pp. 651–655.

[75] N. Bevan, "Usability is Quality of Use", in *Advances in Human Factors/Ergonomics*, vol. 20, Elsevier, 1995, pp. 349–354, ISBN: 978-0-444-81795-2. DOI: 10.1016/S0921-2647(06)80241-8

[76] International Organization for Standardization, "Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and Concepts", International Organization for Standardization, Geneva, Switzerland, Standard ISO 9241-11:2018, 2018.

[77] "Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability", International Organization for Standardization, Geneva, Switzerland, Standard, 1998, Volume: 9241-11:1998. [Online]. Available: https://www.iso.org/standard/16883.html

[78] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research", *International Journal of Human-Computer Studies*, vol. 64, no. 2, pp. 79–102, Feb. 2006. DOI: 10.1016/j.ijhcs.2005.06.002

[79] J. Sauro and J. R. Lewis, "Correlations among prototypical usability metrics: Evidence for the construct of usability", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston MA USA: ACM, Apr. 4, 2009, pp. 1609–1618, ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518947

[80] J. R. Lewis, "Measuring Perceived Usability: SUS, UMUX, and CSUQ Ratings for Four Everyday Products", *International Journal of Human–Computer Interaction*, vol. 35, no. 15, pp. 1404–1419, Sep. 2019. DOI: 10.1080/10447318.2018.1533152

[81] M. Hassenzahl and N. Tractinsky, "User experience - a research agenda", *Behaviour & Information Technology*, vol. 25, no. 2, pp. 91–97, Mar. 2006. DOI: 10.1080/01449290500330331

[82] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: A survey approach", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston MA USA: ACM, Apr. 2009, pp. 719–728, ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518813

[83] International Organization for Standardization, *Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems*, Issue: ISO 9241-210:2019, 2019. [Online]. Available: https://iso.org

[84] M. Hassenzahl, "The Thing and I: Understanding the Relationship Between User and Product", in *Funology*, ser. Human-Computer Interaction Series, M. A. Blythe et al., Eds., vol. 3, Dordrecht: Springer Netherlands, 2003, pp. 31–42, ISBN: 978-1-4020-2966-0. DOI: 10.1007/1-4020-2967-5_4

[85] S. Diefenbach, N. Kolb, and M. Hassenzahl, "The 'hedonic' in human-computer interaction: History, contributions, and future research directions", in *Proceedings of the 2014 conference on Designing interactive systems*, Vancouver BC Canada: ACM, Jun. 2014, pp. 305–314, ISBN: 978-1-4503-2902-6. DOI: 10.1145/2598510.2598549

[86] M. Thüring and S. Mahlke, "Usability, aesthetics and emotions in human–technology interaction", *International Journal of Psychology*, vol. 42, no. 4, pp. 253–264, Aug. 2007. DOI: 10.1080/00207590701396674

[87] H. M. Hassan and G. H. Galal-Edeen, "From usability to user experience", in *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Okinawa: IEEE, Nov. 2017, pp. 216–222, ISBN: 978-1-5090-6664-3. DOI: 10.1109/ICIIBMS.2017.8279761

[88] N. Bevan, "What is the difference between the purpose of usability and user experience evaluation methods?", *Proceedings of the Workshop UXEM*, Jan. 2009.

[89] J. Brooke, "SUS: A "quick and dirty" usability scale", in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, Eds., London: Taylor & Francis, 1996, pp. 189–194, ISBN: 978-0-7484-0460-5.

[90] A. Bangor, P. T. Kortum, and J. T. Miller, "An Empirical Evaluation of the System Usability Scale", *International Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, Jul. 2008. DOI: 10.1080/10447310802205776

[91] J. Sauro and J. R. Lewis, "When designing usability questionnaires, does it hurt to be positive?", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 2215–2224, ISBN: 9781450302289. DOI: 10.1145/1978942.1979266

[92] T. J. Brix, A. Janssen, M. Storck, and J. Varghese, "Comparison of German Translations of the System Usability Scale – Which to Take?", in *German Medical Data Sciences 2023 – Science. Close to People.* Ser. Studies in Health Technology and Informatics, R. Röhrig et al., Eds., IOS Press, Sep. 12, 2023, ISBN: 978-1-64368-428-4. DOI: 10.3233/SHTI230699

[93] B. Rummel, *System Usability Scale – jetzt auch auf Deutsch*, Published: SAP Community Blog, 2015. Accessed: Oct. 20, 2026. [Online]. Available: https://community.sap.com/t5/additional-blog-posts-by-sap/system-usability-scale-jetzt-auch-auf-deutsch/ba-p/13487686

[94] S. A. C. Perrig, N. von Felten, B. Vollenwyder, and K. Opwis, "Development and psychometric validation of a positively worded German version of the System Usability Scale (SUS)", *International Journal of Human–Computer Interaction*, vol. 41, no. 16, pp. 10 399–10 419, 2025. DOI: 10.1080/10447318.2024.2434720

[95] J. Sauro and J. R. Lewis, "Standardized usability questionnaires", in *Quantifying the User Experience*, Elsevier, 2016, pp. 185–248, ISBN: 978-0-12-802308-2. DOI: 10.1016/B978-0-12-802308-2.00008-4

[96] K. Finstad, "The Usability Metric for User Experience", *Interacting with Computers*, vol. 22, no. 5, pp. 323–327, Sep. 2010. DOI: 10.1016/j.intcom.2010.04.004

[97] J. R. Lewis, B. S. Utesch, and D. E. Maher, "UMUX-LITE: When there's no time for the SUS", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris France: ACM, Apr. 2013, pp. 2099–2102, ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2481287

[98] P. T. Kortum and A. Bangor, "Usability Ratings for Everyday Products Measured With the System Usability Scale", *International Journal of Human-Computer Interaction*, vol. 29, no. 2, pp. 67–76, Jan. 2013. DOI: 10.1080/10447318.2012.681221

[99] J. R. Lewis, B. S. Utesch, and D. E. Maher, "Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability", *International Journal of Human–Computer Interaction*, vol. 31, no. 8, pp. 496–505, 2015. DOI: 10.1080/10447318.2015.1064654

[100] J. R. Lewis, "Measuring User Experience With 3, 5, 7, or 11 Points: Does It Matter?", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 63, no. 6, pp. 999–1011, Sep. 2021. DOI: 10.1177/0018720819881312

[101] J. R. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use", *International Journal of Human–Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995. DOI: 10.1080/10447319509526110

[102] F. F. Reichheld, "The One Number You Need to Grow", *Harvard Business Review*, vol. 81, no. 12, pp. 46–54, 124, Dec. 2003. [Online]. Available: https://hbr.org/2003/12/the-one-number-you-need-to-grow

[103] B. Laugwitz, T. Held, and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire", in *HCI and Usability for Education and Work*, ser. Lecture Notes in Computer Science, A. Holzinger, Ed., vol. 5298, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 63–76, ISBN: 978-3-540-89349-3. DOI: 10.1007/978-3-540-89350-9_6

[104] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests", *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951. DOI: 10.1007/BF02310555

[105] A. Vaswani et al., "Attention is All you Need", in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[106] H. Naveed et al., "A Comprehensive Overview of Large Language Models", *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1–72, Oct. 31, 2025. DOI: 10.1145/3744746

[107] J. Hoffmann et al., "Training Compute-Optimal Large Language Models", *arXiv preprint arXiv:2203.15556*, 2022. DOI: 10.48550/ARXIV.2203.15556

[108] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback", in *Advances in Neural Information Processing Systems 35*, New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022, pp. 27 730–27 744, ISBN: 978-1-7138-7108-8. DOI: 10.52202/068431-2011

[109] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model", in *Advances in Neural Information Processing Systems 36*, New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp. 53 728–53 741, ISBN: 978-1-7138-9911-2. DOI: 10.52202/075280-2338

[110] A. Grattafiori et al., "The Llama 3 Herd of Models", *arXiv preprint arXiv:2407.21783*, Nov. 23, 2024. DOI: 10.48550/arXiv.2407.21783

[111] T. B. Brown et al., "Language Models are Few-Shot Learners", *arXiv preprint arXiv:2005.14165*, 2020. DOI: 10.48550/ARXIV.2005.14165

[112] J. Wei et al., "Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models", in *Advances in Neural Information Processing Systems 35*, New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022, pp. 24 824–24 837, ISBN: 978-1-7138-7108-8. DOI: 10.52202/068431-1800

[113] S. Minaee et al., "Large Language Models: A Survey", *arXiv preprint arXiv:2402.06196*, 2024. DOI: 10.48550/ARXIV.2402.06196

[114] Z. Ji et al., "Survey of Hallucination in Natural Language Generation", *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, Dec. 31, 2023. DOI: 10.1145/3571730

[115] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 3, 2021, pp. 610–623, ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922

[116] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", *arXiv preprint arXiv:2005.11401*, 2020. DOI: 10.48550/ARXIV.2005.11401

[117] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey", *arXiv preprint arXiv:2312.10997*, 2023. DOI: 10.48550/ARXIV.2312.10997

[118] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval Augmentation Reduces Hallucination in Conversation", in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 3784–3803. DOI: 10.18653/v1/2021.findings-emnlp.320

[119] N. Jegham, M. Abdelatti, C. Y. Koh, L. Elmoubarki, and A. Hendawi, "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference", *arXiv preprint arXiv:2505.09598*, 2025. DOI: 10.48550/ARXIV.2505.09598

[120] Q. Zhang, Z. Liu, and S. Pan, "The Rise of Small Language Models", *IEEE Intelligent Systems*, vol. 40, no. 1, pp. 30–37, Jan. 2025. DOI: 10.1109/MIS.2024.3517792

[121] P. Belcak et al., "Small Language Models are the Future of Agentic AI", *arXiv preprint arXiv:2506.02153*, Sep. 23, 2026. DOI: 10.48550/ARXIV.2506.02153

[122] Z. Lu et al., "Small Language Models: Survey, Measurements, and Insights", *arXiv preprint arXiv:2409.15790*, Feb. 26, 2025. DOI: 10.48550/arXiv.2409.15790

[123] F. Wang et al., "A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness", *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 6, pp. 1–87, Dec. 31, 2025. DOI: 10.1145/3768165

[124] Project Apertus et al., "Apertus 1.5: From Open Text Models to Open Multimodal Reasoners", *arXiv preprint arXiv:2610.15542*, Oct. 14, 2026. DOI: 10.48550/ARXIV.2610.15542

[125] Swiss AI Initiative. "Apertus: Fully open foundation model for sovereign AI", Accessed: Nov. 7, 2026. [Online]. Available: https://www.apertus-ai.org/

[126] S. Kulviwat, G. C. Bruner II, A. Kumar, S. A. Nasco, and T. Clark, "Toward a unified theory of consumer acceptance technology", *Psychology & Marketing*, vol. 24, no. 12, pp. 1059–1084, Dec. 2007. DOI: 10.1002/mar.20196

[127] L. Mlekus, D. Bentler, A. Paruzel, A.-L. Kato-Beiderwieden, and G. W. Maier, "How to raise technology acceptance: User experience characteristics as technology-inherent determinants", *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, vol. 51, no. 3, pp. 273–283, Sep. 2020. DOI: 10.1007/s11612-020-00529-7

[128] S. Borsci et al., "The Chatbot Usability Scale: The Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents", *Personal and Ubiquitous Computing*, vol. 26, no. 1, pp. 95–119, Feb. 2022. DOI: 10.1007/s00779-021-01582-9

[129] S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. McTear, "Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?", in *Proceedings of the 31st European Conference on Cognitive Ergonomics*, BELFAST United Kingdom: ACM, Sep. 2019, pp. 207–214, ISBN: 978-1-4503-7166-7. DOI: 10.1145/3335082.3335094

[130] S. Borsci, M. Schmettow, A. Malizia, A. Chamberlain, and F. Van Der Velde, "A confirmatory factorial analysis of the Chatbot Usability Scale: A multilanguage validation", *Personal and Ubiquitous Computing*, vol. 27, no. 2, pp. 317–330, Apr. 2023. DOI: 10.1007/s00779-022-01690-0

[131] S. Holmes, R. Bond, A. Moorhead, J. Zheng, V. Coates, and M. McTear, "Towards Validating a Chatbot Usability Scale", in *Design, User Experience, and Usability*, ser. Lecture Notes in Computer Science, A. Marcus, E. Rosenzweig, and M. M. Soares, Eds., vol. 14033, Cham: Springer Nature Switzerland, 2023, pp. 321–339, ISBN: 978-3-031-35707-7. DOI: 10.1007/978-3-031-35708-4_24

[132] S. Borsci and M. Schmettow, "Re-examining the chatBot Usability Scale (BUS-11) to assess user experience with customer relationship management chatbots", *Personal and Ubiquitous Computing*, vol. 28, no. 6, pp. 1033–1044, Dec. 2024. DOI: 10.1007/s00779-024-01834-4

[133] A. Følstad et al., "Future directions for chatbot research: An interdisciplinary research agenda", *Computing*, vol. 103, no. 12, pp. 2915–2942, Dec. 2021. DOI: 10.1007/s00607-021-01016-7

[134] S. Moussawi, M. Koufaris, and R. Benbunan-Fich, "How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents", *Electronic Markets*, vol. 31, no. 2, pp. 343–364, Jun. 2021. DOI: 10.1007/s12525-020-00411-w

[135] T. Fernandes and E. Oliveira, "Understanding consumers' acceptance of automated technologies in service encounters: Drivers of digital voice assistants adoption", *Journal of Business Research*, vol. 122, pp. 180–191, Jan. 2021. DOI: 10.1016/j.jbusres.2020.08.058

[136] A. Strzelecki, "Students' Acceptance of ChatGPT in Higher Education: An Extended Unified Theory of Acceptance and Use of Technology", *Innovative Higher Education*, vol. 49, no. 2, pp. 223–245, Apr. 2024. DOI: 10.1007/s10755-023-09686-1

[137] S. Kelly, S.-A. Kaye, and O. Oviedo-Trespalacios, "What factors contribute to the acceptance of artificial intelligence? A systematic review", *Telematics and Informatics*, vol. 77, p. 101 925, Feb. 2023. DOI: 10.1016/j.tele.2022.101925

[138] J. Weizenbaum, "ELIZA—A Computer Program For the Study of Natural Language Communication Between Man and Machine", *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1, 1966. DOI: 10.1145/365153.365168

[139] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology", in *Artificial Intelligence Applications and Innovations*, ser. IFIP Advances in Information and Communication Technology, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds., vol. 584, Cham: Springer International Publishing, 2020, pp. 373–383, ISBN: 978-3-030-49185-7. DOI: 10.1007/978-3-030-49186-4_31

[140] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications", *Machine Learning with Applications*, vol. 2, p. 100 006, Dec. 2020. DOI: 10.1016/j.mlwa.2020.100006

[141] A. Følstad and P. B. Brandtzæg, "Chatbots and the new world of HCI", *Interactions*, vol. 24, no. 4, pp. 38–42, Jun. 23, 2017. DOI: 10.1145/3085558

[142] P. B. Brandtzaeg and A. Følstad, "Why People Use Chatbots", in *Internet Science*, ser. Lecture Notes in Computer Science, I. Kompatsiaris et al., Eds., vol. 10673, Cham: Springer International Publishing, 2017, pp. 377–392, ISBN: 978-3-319-70283-4. DOI: 10.1007/978-3-319-70284-1_30

[143] S. Moussawi, M. Koufaris, and R. Benbunan-Fich, "The role of user perceptions of intelligence, anthropomorphism, and self-extension on continuance of use of personal intelligent agents", *European Journal of Information Systems*, vol. 32, no. 3, pp. 601–622, May 4, 2023. DOI: 10.1080/0960085X.2021.2018365

[144] S. Ghani, A. Håkansson, O. Pasichnyi, and H. Shahrokni, "Conversational Agents for Building Energy Efficiency - Advising Housing Cooperatives in Stockholm on Reducing Energy Consumption", *IOP Conference Series: Earth and Environmental Science*, vol. 1546, no. 1, p. 012 027, Nov. 1, 2025. DOI: 10.1088/1755-1315/1546/1/012027

[145] L. Shu and D. Zhao, "Can AI Make Energy Retrofit Decisions? An Evaluation of Large Language Models", *arXiv preprint arXiv:2509.06307*, 2025. DOI: 10.48550/ARXIV.2509.06307

[146] A. Hernández-Cano et al., "Apertus: Democratizing Open and Compliant LLMs for Global Language Environments", in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, California, United States: Association for Computational Linguistics, 2026, pp. 46 877–46 955. DOI: 10.18653/v1/2026.acl-long.2172

[147] European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, Official Journal of the European Union, L, 2024/1689, 12 July 2024, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj

[148] Hugging Face. "Inference providers", Accessed: Nov. 4, 2026. [Online]. Available: https://huggingface.co/docs/inference-providers/index

[149] Public AI. "Public AI inference utility", Accessed: Nov. 4, 2026. [Online]. Available: https://publicai.co/

[150] H. Inan et al., "Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations", *arXiv preprint arXiv:2312.06674*, 2023. DOI: 10.48550/ARXIV.2312.06674

[151] T. Rebedea, R. Dinu, M. N. Sreedhar, C. Parisien, and J. Cohen, "NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Y. Feng and E. Lefever, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 431–445. DOI: 10.18653/v1/2023.emnlp-demo.40

[152] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365

[153] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550

[154] E. Williams, "Experimental Designs Balanced for the Estimation of Residual Effects of Treatments", *Australian Journal of Scientific Research Series A: Physical Sciences*, vol. 2, no. 2, pp. 149–168, Jun. 1, 1949. DOI: 10.1071/CH9490149

[155] K. Caine, "Local Standards for Sample Size at CHI", in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose California USA: ACM, May 7, 2016, pp. 981–992, ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858498

[156] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences", *Behavior Research Methods*, vol. 39, no. 2, pp. 175–191, May 2007. DOI: 10.3758/BF03193146

[157] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses", *Behavior Research Methods*, vol. 41, no. 4, pp. 1149–1160, Nov. 2009. DOI: 10.3758/BRM.41.4.1149

[158] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988, ISBN: 978-0-8058-0283-2.

[159] D. Harborth and S. Pape, "German Translation of the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) Questionnaire", *SSRN Electronic Journal*, 2018. DOI: 10.2139/ssrn.3147708

[160] Artificial Intelligence Center Hamburg (ARIC) e.V. "ARIC – Artificial Intelligence Center Hamburg", Accessed: Nov. 5, 2026. [Online]. Available: https://aric-hamburg.de/

[161] AM Crypto, *VeraCrypt – free open source disk encryption*, version 1.26.31, 2026. Accessed: Nov. 5, 2026. [Online]. Available: https://veracrypt.io/en/Home.html

[162] The pandas development team, *Pandas-dev/pandas: Pandas*, version v3.0.3, May 2026. DOI: 10.5281/zenodo.20127038

[163] W. McKinney, "Data structures for statistical computing in Python", in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a

[164] T. E. Oliphant, *A Guide to NumPy*. USA: Trelgol Publishing, 2006.

[165] C. R. Harris et al., "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2

[166] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python", *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2

[167] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python", in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011

[168] J. D. Hunter, "Matplotlib: A 2D graphics environment", *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55

[169] The Apache Arrow Developers, *PyArrow: Python library for Apache Arrow*, version 25.0.1, 2026. Accessed: Nov. 5, 2026. [Online]. Available: https://arrow.apache.org/

[170] S. Holm, "A simple sequentially rejective multiple test procedure", *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: https://www.jstor.org/stable/4615733

[171] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)", *Biometrika*, vol. 52, no. 3–4, pp. 591–611, Dec. 1, 1965. DOI: 10.1093/biomet/52.3-4.591

[172] H. Levene, "Robust tests for equality of variances", in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, Eds., Palo Alto, CA: Stanford University Press, 1960, pp. 278–292.

[173] R. D. Cook, "Detection of Influential Observation in Linear Regression", *Technometrics*, vol. 19, no. 1, pp. 15–18, Feb. 1977. DOI: 10.1080/00401706.1977.10489493

[174] X.-l. Meng, R. Rosenthal, and D. B. Rubin, "Comparing correlated correlation coefficients", *Psychological Bulletin*, vol. 111, no. 1, pp. 172–175, Jan. 1992. DOI: 10.1037/0033-2909.111.1.172

[175] G. Y. Zou, "Toward using confidence intervals to compare correlations", *Psychological Methods*, vol. 12, no. 4, pp. 399–413, Dec. 2007. DOI: 10.1037/1082-989X.12.4.399

[176] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, pp. 289–300, Jan. 1, 1995. DOI: 10.1111/j.2517-6161.1995.tb02031.x

[177] J. R. Lewis and J. Sauro, "Item benchmarks for the system usability scale", *Journal of Usability Studies*, vol. 13, no. 3, pp. 158–167, May 2018.

[178] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R^2 from generalized linear mixed-effects models", *Methods in Ecology and Evolution*, vol. 4, no. 2, R. B. O'Hara, Ed., pp. 133–142, Feb. 2013. DOI: 10.1111/j.2041-210x.2012.00261.x

[179] M. T. Orne, "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications", *American Psychologist*, vol. 17, no. 11, pp. 776–783, Nov. 1962. DOI: 10.1037/h0043424

[180] A. J. Fridlund, "Sociality of solitary smiling: Potentiation by an implicit audience", *Journal of Personality and Social Psychology*, vol. 60, no. 2, pp. 229–240, Feb. 1991. DOI: 10.1037/0022-3514.60.2.229

[181] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation", in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, St. Julians, Malta: Association for Computational Linguistics, 2024, pp. 150–158. DOI: 10.18653/v1/2024.eacl-demo.16

[182] A. Bangor, P. Kortum, and J. Miller, "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale", *Journal of Usability Studies*, vol. 4, pp. 114–123, Apr. 2009.

[183] Bundesgerichtshof, *Urteil vom 18.07.2007 – VIII ZR 285/06*, NJW 2007, 3122, Jahresfrist nach § 558 Abs. 1 S. 2 BGB; Vorinstanzen LG Görlitz, AG Görlitz, Jul. 18, 2007.

[184] Bundesgerichtshof, *Urteil vom 17.12.2014 – VIII ZR 87/13*, Mieterhöhung nach Modernisierung in mehreren Bauabschnitten; Vorinstanz LG Berlin, 65 S 207/12, Dec. 17, 2014.

[185] Bundesgerichtshof, *Urteil vom 17.12.2014 – VIII ZR 88/13*, NJW 2015, 934; NZM 2015, 198; WuM 2015, 165, Mieterhöhung nach Modernisierung in mehreren Bauabschnitten; Darlegungslast, ersparte Instandhaltungskosten, Dec. 17, 2014.

[186] Bundesgerichtshof, *Urteil vom 17.12.2014 – VIII ZR 89/13*, Parallelverfahren zu VIII ZR 87/13 und 88/13; Vorinstanz LG Berlin, 65 S 410/11, Dec. 17, 2014.

[187] Bundesgerichtshof, *Versäumnisurteil vom 17.06.2020 – VIII ZR 81/19*, NZM 2020, 795, Teilunwirksame Modernisierungsmieterhöhung; Abzug des Instandhaltungsanteils auch bei nicht fälliger Instandsetzung; Beweislastverteilung, Jun. 17, 2020.

[188] Bundesgerichtshof, *Urteil vom 20.07.2022 – VIII ZR 361/21*, NZM 2022, 795; WuM 2022, 542; ZMR 2022, 951, Formelle Anforderungen an die Modernisierungsmieterhöhung (§ 559b BGB); Angabe ersparter Erhaltungskosten, Jul. 20, 2022.

[189] Bundesgerichtshof, *Urteil vom 23.11.2022 – VIII ZR 59/21*, NJW 2023, 360; NZM 2023, 213; ZMR 2023, 351, Formelle Anforderungen an die Mieterhöhungserklärung nach Modernisierung (§ 559b BGB), Nov. 23, 2022.

[190] Bundesgerichtshof, *Beschluss vom 21.02.2023 – VIII ZR 106/21*, openJur, Hinweisbeschluss nach § 552a ZPO; formelle Anforderungen an die Modernisierungsmieterhöhung, Feb. 21, 2023.

[191] U. Börstinghaus, "Die aktuelle Rechtsprechung des BGH zum Mietrecht", *Neue Zeitschrift für Miet- und Wohnungsrecht (NZM)*, pp. 521–538, 2021.

[192] B. Schindler, "Modernisierungsmieterhöhung und Instandsetzungskosten. Besprechung von BGH NZM 2020, 795", *Neue Zeitschrift für Miet- und Wohnungsrecht (NZM)*, pp. 258–264, 2021.

[193] M. Artz, "§ 559 BGB – Mieterhöhung nach Modernisierungsmaßnahmen", in *Münchener Kommentar zum Bürgerlichen Gesetzbuch. Band 5: Schuldrecht – Besonderer Teil II, §§ 535–630h*, F. J. Säcker, R. Rixecker, H. Oetker, and B. Limperg, Eds., 9th ed., vol. 5, München: C.H. Beck, 2023.

A Appendix

A Appendix **109**
A.1 Demographic Questionnaire and Response Frequencies 110
A.2 System Usability Scale 112
A.3 System Usability Scale – German translation and response frequencies . . . 113
A.4 UTAUT Questionnaire 114
A.5 UTAUT questionnaire – German translation and response frequencies . . . 116
A.6 Ethics Committee Approval 118
A.7 Model Prompts 119
A.7.1 System Prompt of Application A: Guided Assistant 119
A.7.2 System Prompt of Application B: Basic Chat 121
A.8 RAG Documents 123
A.9 Additional Results 124

A.1 Demographic Questionnaire and Response Frequencies

Table A.1: Demographic questionnaire with the number of participants who selected each option ($N = 24$)

Nr.	Frage	Antwortoptionen	*n*
1	Welchem Geschlecht ordnen Sie sich zu?	(1) Männlich	14
		(2) Weiblich	10
		(3) Divers	0
		(4) Keine Angabe	0
2	Wie alt sind Sie (in Jahren)?	(1) 18–24	4
		(2) 25–34	12
		(3) 35–44	1
		(4) 45–54	3
		(5) 55–64	2
		(6) 65 Jahre oder älter	2
3	Wie schätzen Sie Ihre Deutschkenntnisse ein?	(1) Muttersprache	21
		(2) Fließend	3
		(3) Fortgeschritten	0
		(4) Grundkenntnisse	0
4	Was ist Ihr höchster Bildungsabschluss?	(1) Kein Abschluss	0
		(2) Hauptschulabschluss	1
		(3) Mittlere Reife	0
		(4) Berufsausbildung	0
		(5) Fachhochschulreife / Abitur	14
		(6) Bachelor	2
		(7) Master	3
		(8) Diplom	3
		(9) Staatsexamen	0
		(10) Promotion	0
		(11) Sonstiges (bitte angeben)[a]	1

Continuation on the next page

Table A.1 – *Continuation*

Nr.	Frage	Antwortoptionen	*n*
5	Wie schätzen Sie Ihre Affinität im Umgang mit digitalen Anwendungen ein?	(1) Sehr gering	0
		(2) Gering	1
		(3) Mittel	4
		(4) Hoch	11
		(5) Sehr hoch	8
6	Wie häufig nutzen Sie digitale Assistenten oder Chatbots wie ChatGPT, Perplexity oder Claude?	(1) Täglich	11
		(2) Wöchentlich	6
		(3) Monatlich	4
		(4) Seltener	2
		(5) Nie	1
7	Wie schätzen Sie Ihr Wissen in dem Bereich Mietrecht / Immobilienrecht ein?	(1) Sehr gering	10
		(2) Gering	8
		(3) Mittel	6
		(4) Hoch	0
		(5) Sehr hoch	0

[a] Free-text answer: "Abitur und Vorbereitungsdienst (quasi Ausbildung zur Beamtin)".

A.2 System Usability Scale

Table A.2: System Usability Scale (SUS) items [89]

Item	Statement
SUS1	I think that I would like to use this system frequently.
SUS2	I found the system unnecessarily complex.
SUS3	I thought the system was easy to use.
SUS4	I think that I would need the support of a technical person to be able to use this system.
SUS5	I found the various functions in this system were well integrated.
SUS6	I thought there was too much inconsistency in this system.
SUS7	I would imagine that most people would learn to use this system very quickly.
SUS8	I found the system very cumbersome to use.
SUS9	I felt very confident using the system.
SUS10	I needed to learn a lot of things before I could get going with this system.

A.3 System Usability Scale – German translation and response frequencies

Table A.3: System Usability Scale (SUS) – German translation [93], with the number of participants who selected each response option per application ($n = 24$ each; 1 = *Stimme gar nicht zu* to 5 = *Stimme voll zu*). Raw responses before reverse coding; the most frequent option is printed in bold.

		Application A					**Application B**				
ID	**Frage**	1	2	3	4	5	1	2	3	4	5
SUS-1	Ich denke, dass ich die Plattform häufig nutzen möchte.	1	3	4	7	**9**	3	2	3	**9**	7
SUS-2	Ich empfand die Plattform als unnötig komplex.	**14**	7	0	2	1	**8**	**8**	0	7	1
SUS-3	Ich glaube, dass die Plattform leicht zu verwenden war.	1	1	1	7	**14**	2	4	3	**11**	4
SUS-4	Ich denke, dass ich die Unterstützung einer technisch versierten Person benötigen würde, um die Plattform nutzen zu können.	**14**	5	2	3	0	**10**	6	2	2	4
SUS-5	Ich fand, dass die verschiedenen Funktionen in der Plattform gut integriert waren.	0	2	6	**8**	**8**	1	1	6	**8**	**8**
SUS-6	Ich glaube, dass es zu viele Inkonsistenzen in der Plattform gab.	**11**	8	5	0	0	**15**	3	5	0	1
SUS-7	Ich könnte mir vorstellen, dass die meisten Menschen sehr schnell lernen würden, die Plattform zu benutzen.	1	0	1	6	**16**	2	2	1	**10**	9
SUS-8	Ich fand die Plattform sehr umständlich in der Benutzung.	**14**	5	3	1	1	**10**	6	1	5	2
SUS-9	Ich fühlte mich im Umgang mit der Plattform sehr sicher.	0	2	2	9	**11**	4	3	**7**	6	4
SUS-10	Ich musste eine Menge Dinge lernen, bevor ich mit der Plattform loslegen konnte.	**15**	7	2	0	0	**8**	3	5	4	4

A.4 UTAUT Questionnaire

Table A.4: Items used in estimating UTAUT (adapted from [11])

Item	Statement
Performance expectancy	
U6	I would find the system useful in my job.
RA1	Using the system enables me to accomplish tasks more quickly.
RA5	Using the system increases my productivity.
OE7	If I use the system, I will increase my chances of getting a raise.
Effort expectancy	
EOU3	My interaction with the system would be clear and understandable.
EOU5	It would be easy for me to become skillful at using the system.
EOU6	I would find the system easy to use.
EU4	Learning to operate the system is easy for me.
Attitude toward using technology	
A1	Using the system is a bad/good idea.
AF1	The system makes work more interesting.
AF2	Working with the system is fun.
Affect1	I like working with the system.
Social influence	
SN1	People who influence my behavior think that I should use the system.
SN2	People who are important to me think that I should use the system.
SF2	The senior management of this business has been helpful in the use of the system.
SF4	In general, the organization has supported the use of the system.
Facilitating conditions	
PBC2	I have the resources necessary to use the system.
PBC3	I have the knowledge necessary to use the system.
PBC5	The system is not compatible with other systems I use.
FC3	A specific person (or group) is available for assistance with system difficulties.
Self-efficacy	
	I could complete a job or task using the system…
SE1	If there was no one around to tell me what to do as I go.
SE4	If I could call someone for help if I got stuck.
SE6	If I had a lot of time to complete the job for which the software was provided.
SE7	If I had just the built-in help facility for assistance.
Anxiety	
ANX1	I feel apprehensive about using the system.
ANX2	It scares me to think that I could lose a lot of information using the system by hitting the wrong key.
ANX3	I hesitate to use the system for fear of making mistakes I cannot correct.
ANX4	The system is somewhat intimidating to me.
Behavioral intention to use the system	
BI1	I intend to use the system in the next $\langle n \rangle$ months.

BI2	I predict I would use the system in the next ⟨n⟩ months.
BI3	I plan to use the system in the next ⟨n⟩ months.

A.5 UTAUT questionnaire – German translation and response frequencies

Table A.5: UTAUT-Questionnaire – German translation [159], with the number of participants who selected each response option per application ($n = 24$ each; 1 = *Stimme gar nicht zu* to 7 = *Stimme voll zu*). PE = Performance Expectancy, EE = Effort Expectancy, SI = Social Influence, FC = Facilitating Conditions, BI = Behavioral Intention. Raw responses before reverse coding of UTAUT-13; the most frequent option is printed in bold.

			Application A							**Application B**						
ID	**Kon.**	**Frage**	1	2	3	4	5	6	7	1	2	3	4	5	6	7
UTAUT-1	PE	Ich würde die Plattform in meiner Arbeit als nützlich empfinden.	1	3	1	0	3	**8**	**8**	2	1	3	1	1	7	**9**
UTAUT-2	PE	Die Nutzung der Plattform ermöglicht es mir, Aufgaben schneller zu erledigen.	1	1	2	0	1	8	**11**	3	0	0	3	3	5	**10**
UTAUT-3	PE	Die Nutzung der Plattform steigert meine Produktivität.	1	1	2	1	3	7	**9**	3	0	3	1	3	**7**	**7**
UTAUT-4	PE	Wenn ich die Plattform nutze, erhöhen sich meine Chancen auf eine Gehaltserhöhung.	3	2	2	**15**	1	1	0	3	2	4	**10**	4	1	0
UTAUT-5	EE	Meine Interaktion mit der Plattform wäre klar und verständlich.	0	0	1	1	7	6	**9**	1	2	3	1	3	**7**	**7**
UTAUT-6	EE	Es wäre für mich leicht, den Umgang mit der Plattform zu erlernen.	0	0	1	0	3	6	**14**	1	2	1	0	4	3	**13**
UTAUT-7	EE	Ich würde die Plattform als einfach zu bedienen empfinden.	0	0	1	1	2	7	**13**	2	1	1	0	3	8	**9**
UTAUT-8	EE	Das Erlernen der Bedienung der Plattform fällt mir leicht.	0	0	1	0	1	6	**16**	1	2	1	0	4	5	**11**
UTAUT-9	SI	Personen, die mein Verhalten beeinflussen, denken, dass ich die Plattform nutzen sollte.	3	2	1	5	**6**	4	3	4	2	2	5	**6**	1	4
UTAUT-10	SI	Personen, die mir wichtig sind, denken, dass ich die Plattform nutzen sollte.	3	2	2	4	**8**	2	3	4	3	2	**6**	3	2	4
UTAUT-11	FC	Ich verfüge über die notwendigen Ressourcen, um die Plattform zu nutzen.	0	0	0	2	3	**10**	9	2	0	2	2	3	4	**11**
UTAUT-12	FC	Ich verfüge über das notwendige Wissen, um die Plattform zu nutzen.	0	0	2	1	5	6	**10**	2	1	3	0	4	5	**9**
UTAUT-13	FC	Die Plattform ist nicht mit anderen Systemen kompatibel, die ich nutze.	**11**	4	2	3	1	2	1	**9**	6	1	6	0	1	1

Continuation on the next page

Table A.5 – *Continuation*

			Application A							**Application B**						
ID	**Kon.**	**Frage**	1	2	3	4	5	6	7	1	2	3	4	5	6	7
UTAUT-14	FC	Es steht eine bestimmte Person (oder Gruppe) zur Verfügung, die bei Problemen mit der Plattform Hilfe leisten kann.	1	1	2	**7**	5	3	5	3	1	2	**5**	4	4	**5**
UTAUT-15	BI	Ich beabsichtige, die Plattform in den nächsten Monaten zu nutzen.	1	3	1	3	4	**7**	5	3	4	0	1	4	**6**	**6**
UTAUT-16	BI	Ich gehe davon aus, dass ich die Plattform in den nächsten Monaten nutzen werde.	1	3	1	1	3	**11**	4	3	2	1	1	5	**6**	**6**
UTAUT-17	BI	Ich plane, die Plattform in den nächsten Monaten zu nutzen.	2	2	1	2	3	**9**	5	3	3	1	2	1	**8**	6

A.6 Ethics Committee Approval



HAW HAMBURG

Herr
Kilian Hildebrandt

ETHIKKOMMISSION

DR. CHRISTIANE STANGE
Geschäftsstelle
Ethikkommission

Ihr Antrag auf Stellungnahme

Hamburg, 14.04.2026

Antrag 2026-04

„Evaluating responsible legal Virtual Assistants for increasing user experience"

T +49 40 428 75 9036

ethikkommission@haw-hamburg.de

vom 06.02.2026 in der Überarbeitung vom 30.03.2026

HOCHSCHULE F ANGEWANDTE WISSENSCHAFTEN HAMBURG
Forschung und Transfer
Steindamm 103
20099 Hamburg

HAW-HAMBURG.DE

Sehr geehrter Herr Hildebrandt,

das o.g. Vorhaben wurde nach Prüfung durch die Ethikkommission der HAW Hamburg als grundsätzlich „ethisch unbedenklich" bewertet.

Mit freundlichen Grüßen

[redacted]

Dr. Christiane Stange

(Geschäftsstelle Ethikkommission)

1 / 1

Figure A.1: Approval of the ethics committee of the Hamburg University of Applied Sciences

A.7 Model Prompts

A.7.1 System Prompt of Application A: Guided Assistant

```
Du bist ein freundlicher Fachberater für Modernisierungsmaßnahmen an Gebäudefassaden.
Der Benutzer ist Vermieter/Eigentümer und möchte eine Fassadendämmung (WDVS) durchführen.
Du hilfst bei der technischen und wirtschaftlichen Bewertung – dem sogenannten Funktionalen
 ↪ Kostensplitting.

== ABLAUF ==

Das Gespräch läuft in 6 Phasen ab:

1. EINFÜHRUNG & ADRESSE: Begrüße den Vermieter, stelle dich als Fachberater vor und erkläre:
 ↪ „Ich helfe Ihnen, eine passende Dämmmaßnahme zu finden und die daraus resultierenden
 ↪ Kosten für Sie und Ihre Mieter zu berechnen." Frage direkt nach der Adresse (Straße,
 ↪ Hausnummer, PLZ, Ort).
2. GEBÄUDEDATEN: Erfasse Bezeichnung, Baujahr, Bauart, Energieträger, Gesamtmietfläche,
 ↪ Mietfläche der Wohnung.
3. GEBÄUDEZUSTAND (IST-BEWERTUNG): Erfasse den aktuellen Gebäudezustand (Standsicherheit,
 ↪ Bauphysik, Feuchteschutz, Wärmeschutz, Schallschutz, Brandschutz, Tageslicht, Ästhetik,
 ↪ Wirtschaftlichkeit – je 0–100 %). Erkläre, dass diese Werte den Unterlagen entnommen
 ↪ werden können.
4. PRIORISIERUNG: Bitte den Vermieter, Ästhetik, Umweltschutz und Kosten nach Wichtigkeit zu
 ↪ ordnen (das wichtigste zuerst). Basierend darauf werden passende Maßnahmen vorgeschlagen.
5. MASSNAHMENEMPFEHLUNG: Präsentiere die Top-3-Empfehlungen und lasse den Vermieter
 ↪ auswählen. Nach der Auswahl werden die Kosten berechnet.
6. ERGEBNIS: Zeige die berechneten Kosten und Mieterhöhungen. Frage, ob noch Fragen bestehen.

In Phase 1 stellst du die Adressfragen als Fließtext (nicht nummeriert).
In den Phasen 2–3 stellst du ALLE fehlenden Fragen des Abschnitts auf einmal als nummerierte
 ↪ Liste.
Gehe erst zur nächsten Phase über, wenn alle Felder der aktuellen vollständig beantwortet
 ↪ sind.
Frage NIEMALS nach Werten, die bereits gesammelt wurden.

Am Ende JEDER Antwort in Phase 1–3 schreibe auf einer eigenen Zeile: FELDER: key1,key2,...
Liste dabei nur die Keys der Felder auf, nach denen du gerade gefragt hast.

== STRIKTE REGELN ==

- Begrüße den Benutzer NUR in Phase 1 – danach keine Begrüßung mehr
```

- In Phase 1: Stelle die Adressfragen als Fließtext, nicht als nummerierte Liste
- In Phasen 2–3: Stelle ALLE fehlenden Fragen als nummerierte Liste
- Frage NIEMALS nach Werten, die bereits gesammelt wurden
- Bestätige kurz, wenn alle Felder eines Abschnitts vollständig erfasst wurden, bevor du zum nächsten übergehst
- Schreibe am Ende JEDER Antwort in Phase 1–3: FELDER: key1,key2,...
- Verwende 'Sie' als Anrede
- Vermeide technische Begriffe wie 'Key', 'Parameter', 'Feld'
- Halte deine Antworten kurz und präzise
- Verwende KEINE Briefformalien: keine Anrede wie „Sehr geehrter…", keine Abschlussformel wie „Beste Grüße" oder „Mit freundlichen Grüßen", keinen Platzhalter wie „[Ihr Name]" oder „[Name]"
- Nenne KEINEN eigenen Namen und unterschreibe NICHT mit einem Namen
- Verwende NIEMALS Markdown Formatierungen

== FESTE PROMPT-VORLAGEN ==

Die folgenden Vorlagen werden vom System verwendet. Halte dich exakt an das geforderte Format.

--- KLASSIFIKATION ---
Wenn du eine Benutzereingabe klassifizieren sollst:
Antworte mit GENAU einem Wort: FRAGE oder ANTWORT.
- FRAGE: Der Benutzer stellt eine Rückfrage oder bittet um Erklärung.
- ANTWORT: Der Benutzer liefert einen oder mehrere Werte.

--- EXTRAKTION ---
Wenn du Werte aus einer Benutzereingabe extrahieren sollst:
Antworte NUR mit gültigem JSON. Fehlende Werte weglassen. Zahlen als Zahlen (kein String).

--- KLARSTELLUNG ---
Wenn der Benutzer eine Rückfrage stellt:
Beantworte die Frage präzise in 1–2 Sätzen und stelle anschließend alle noch offenen Fragen des aktuellen Abschnitts erneut.
Schreibe auch dann am Ende: FELDER: key1,key2,...

--- PRIORISIERUNG ---
Wenn du nach der Priorisierung fragst:
Bitte den Vermieter, Ästhetik, Umweltschutz und Kosten nach Wichtigkeit zu ordnen (das wichtigste zuerst).
Stelle dann basierend auf der Priorisierung geeignete Maßnahmen vor.

--- EMPFEHLUNG ---

```
Wenn du Maßnahmen empfiehlst:
Präsentiere die Top 3 als nummerierte Liste mit Name, Preis pro m² und einer kurzen Erklärung
↪ (1–2 Sätze), warum diese Maßnahme zur Priorisierung des Vermieters passt.
Frage danach: „Mit welcher der Maßnahmen möchten Sie fortfahren?"

== KONTEXT ==

- IST = aktueller Zustand vor der Modernisierung
- MOD = erwarteter Zustand nach der Modernisierung
- Prozentwerte: 0 % = sehr schlecht / nicht vorhanden, 100 % = perfekt
- WDVS = Wärmedämmverbundsystem (Fassadendämmung)
- Mietfläche = nutzbare Wohnfläche nach Wohnflächenverordnung (ohne Keller, Dachboden,
↪ Stellplätze)
- Ästhetik = äußeres Erscheinungsbild nach der Modernisierung
- Umweltschutz = ökologische Nachhaltigkeit der Maßnahme
- Kosten = wirtschaftliche Effizienz, möglichst kostengünstig

== RÜCKFRAGEN ==

Wenn der Benutzer fragt statt einen Wert zu liefern, beantworte die Frage präzise in 1–2
↪ Sätzen und stelle anschließend alle noch offenen Fragen des aktuellen Abschnitts erneut.
Schreibe auch dann am Ende: FELDER: key1,key2,...
```

A.7.2 System Prompt of Application B: Basic Chat

```
Du bist ein freundlicher Fachberater für Gebäudemodernisierung und Fassadendämmung.
Du beantwortest Fragen rund um die energetische Sanierung von Wohngebäuden, das Funktionale
↪ Kostensplitting, Wärmedämmverbundsysteme (WDVS) und verwandte Themen.

== DEINE AUFGABE ==

Du hilfst Nutzern, Begriffe, Verfahren und Zusammenhänge im Bereich der Gebäudemodernisierung
↪ zu verstehen. Du erklärst fachliche Konzepte verständlich und beantwortest Fragen auf
↪ Basis deines Wissens und der verfügbaren Fachdokumente.

Du füllst KEINE Formulare aus und sammelst KEINE Daten. Du führst KEINE Berechnungen durch.
↪ Du bist ein reiner Informationsassistent.

Gib immer die Quellen aus dem Kontext mit an, wenn du extra Kontext aus dem RAG verwendet
↪ hast! Nicht den Dateinamen sondern eine APA style citation für die jeweilige Quelle.
```

```
== REGELN ==

- Verwende 'Sie' als Anrede
- Halte deine Antworten klar und verständlich
- Vermeide unnötige Fachbegriffe; wenn nötig, erkläre sie
- Wenn du Informationen aus den bereitgestellten Dokumenten verwendest, verweise auf die
 ↪  Quelle (Verfasser, Jahr und Seitenzahl)
- Verwende bei der Quellenangabe keine Dateiendungen
- Prüfe für die Quelle wenn nötig die entsprechenden Angaben im Dokument selbst
- Wenn du eine Frage nicht beantworten kannst, sage das offen
- Stelle keine Rückfragen, es sei denn, die Frage des Nutzers ist unklar
- Antworte auf Deutsch
- Verwende NIEMALS Markdown Formatierungen

== KONTEXT ==

- WDVS = Wärmedämmverbundsystem (Fassadendämmung)
- Funktionales Kostensplitting = Verfahren zur Aufteilung der Modernisierungskosten zwischen
 ↪  Vermieter und Mieter
- IST = aktueller Zustand vor der Modernisierung
- MOD = erwarteter Zustand nach der Modernisierung
- GEG = Gebäudeenergiegesetz
- BEG = Bundesförderung für effiziente Gebäude
```

A.8 RAG Documents

Table A.6: Documents of the RAG knowledge base (13 German PDF documents, ~480 pages). Documents were selected within the KIMM and IntelMOD projects [7], [8].

#	Document	Subject
BGH rulings on rent increases (8 documents, 101 pages)		
1	18.07.2007, VIII ZR 285/06 [183]	One-year waiting period under § 558(1) sentence 2 BGB
2	17.12.2014, VIII ZR 87/13 [184]	Rent increase after modernization in several construction phases
3	17.12.2014, VIII ZR 88/13 [185]	Modernization in several construction phases; burden of presentation; saved maintenance costs
4	17.12.2014, VIII ZR 89/13 [186]	Parallel proceedings to VIII ZR 87/13 and 88/13
5	17.06.2020, VIII ZR 81/19 (default judgment) [187]	Partially invalid modernization rent increase; deduction of maintenance share even if repairs not yet due; burden of proof
6	20.07.2022, VIII ZR 361/21 [188]	Formal requirements of the modernization rent increase (§ 559b BGB); statement of saved maintenance costs
7	23.11.2022, VIII ZR 59/21 [189]	Formal requirements of the rent increase declaration after modernization (§ 559b BGB)
8	21.02.2023, VIII ZR 106/21 (order) [190]	Notice order under § 552a ZPO; formal requirements of the modernization rent increase
Law journal articles (2 documents, 43 pages)		
9	Börstinghaus [191], NZM 2021, 521–538	Review of current BGH case law on tenancy law
10	Schindler [192], NZM 2021, 258–264	Modernization rent increase and repair costs; case note on BGH NZM 2020, 795 (VIII ZR 81/19)
Case law overview (1 document, 34 pages)		
11	beck-online overview “Die aktuelle Rechtsprechung des BGH zum Mietrecht”	Overview of BGH decisions on tenancy law
Commentary (1 document, 2 pages)		
12	Artz [193], MüKo BGB, 9th ed., § 559	Commentary on rent increases after modernization measures (§ 559 BGB)
Dissertation (1 document, 300 pages)		
13	David [9], HafenCity Universität Hamburg	Functional cost splitting for determining rent increases after energy-efficiency measures

A.9 Additional Results

Table A.8: Hypothesis models pooled over both applications (linear mixed-effects models)

Model	Predictor	b	95 % CI	β	p
H1a (PE)	SUS	0.047	[0.033, 0.061]	0.69	$< .001$
H1b (EE)	SUS	0.059	[0.049, 0.068]	0.91	$< .001$
H2a (PE)	Valence	−0.019	[−0.095, 0.057]	−0.13	.619
	Arousal	0.024	[−0.014, 0.063]	0.23	.218
	Valence × arousal	−0.001	[−0.005, 0.002]	−0.16	.433
H2b (EE)	Valence	−0.006	[−0.074, 0.063]	−0.04	.873
	Arousal	0.001	[−0.035, 0.037]	0.01	.969
	Valence × arousal	−0.003	[−0.007, 0.001]	−0.31	.148
H2c (SI)	Valence	−0.011	[−0.096, 0.074]	−0.06	.801
	Arousal	0.017	[−0.023, 0.058]	0.14	.395
	Valence × arousal	−0.001	[−0.005, 0.003]	−0.10	.539

Note. All 48 sessions; random intercept per participant; application, model size and session position as covariates (not shown). Wald tests.

Table A.9: Single-emotion mixed models for the UTAUT constructs

	PE		EE		SI		BI	
Emotion	b	χ^2 (p)	b	χ^2 (p)	b	χ^2 (p)	b	χ^2 (p)
Anger	0.32	1.55 (.213)	0.43	3.30 (.069)	0.31	1.30 (.255)	0.45	1.85 (.174)
Contempt	0.14	0.37 (.543)	0.00	0.00 (.981)	0.35	2.77 (.096)	0.03	0.01 (.909)
Disgust	−0.01	0.00 (.965)	0.09	0.16 (.693)	0.06	0.05 (.819)	−0.20	0.35 (.551)
Fear	0.47	2.96 (.086)	0.39	2.51 (.113)	0.33	1.06 (.304)	0.61	3.20 (.074)
Joy	−0.04	0.03 (.866)	0.14	0.40 (.526)	−0.16	0.43 (.510)	−0.32	1.01 (.316)
Sadness	0.34	2.32 (.128)	0.24	1.27 (.261)	0.17	0.48 (.488)	0.58	4.13 (.042)
Surprise	0.41	2.17 (.141)	0.19	0.58 (.447)	0.11	0.11 (.740)	0.43	1.50 (.220)

Note. All 48 sessions; one model per emotion and construct with the log-transformed, standardized emotion evidence, application, model size, and sequence as fixed effects and a random intercept per participant. b = change in the UTAUT score (1–7) per *SD* of the log emotion evidence (REML). $\chi^2(1)$ and uncorrected p from likelihood-ratio tests; after Holm correction across the seven emotions per construct, all $p \geq .295$.

Erklärung zur selbständigen Bearbeitung einer Abschlussarbeit

Gemäß der Allgemeinen Prüfungs- und Studienordnung ist zusammen mit der Abschlussarbeit eine schriftliche Erklärung abzugeben, in der der Studierende bestätigt, dass die Abschlussarbeit „-- bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit [(§ 18 Abs. 1 APSO-TI-BM bzw. § 21 Abs. 1 APSO-INGI)] -- ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt wurden. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich zu machen."

Quelle: § 16 Abs. 5 APSO-TI-BM bzw. § 15 Abs. 6 APSO-INGI

Dieses Blatt, mit der folgenden Erklärung, ist nach Fertigstellung der Abschlussarbeit durch den Studierenden auszufüllen und jeweils mit Originalunterschrift als letztes Blatt in das Prüfungsexemplar der Abschlussarbeit einzubinden.
Eine unrichtig abgegebene Erklärung kann -auch nachträglich- zur Ungültigkeit des Studienabschlusses führen.

Erklärung zur selbständigen Bearbeitung der Arbeit

Hiermit versichere ich,

Name: Hildebrandt

Vorname: Kilian

dass ich die vorliegende Masterarbeit mit dem Thema:
Acceptance of AI Virtual Assistants in the Context of Housing Modernization
ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht. Zur sprachlichen Überarbeitung, zur Prüfung von Zitaten und Querverweisen sowie zur Unterstützung bei der Programmierung der Auswertungsskripte und der Anwendungen wurden KI-Werkzeuge (Claude von Anthropic) eingesetzt. Konzeption, Durchführung, Auswertung und Interpretation der Studie sowie alle inhaltlichen Aussagen stammen von mir; alle Ergebnisse wurden von mir geprüft.

Hamburg	17.11.2026	
Ort	Datum	Unterschrift im Original