

# **Vertrauen in Künstliche Intelligenz**

**Wie Framing das Vertrauen in LLM-basierte Applikationen- und Antworten**

**beeinflusst**

## **Vorstudie Bachelorarbeit**

Fabian Ryf

Hochschule Luzern, Wirtschaft

HSLU-W

BSc Business Psychology (BP)

Markt- und Konsumentenpsychologie

Dr. Andreas Hüsser

05.12.2025

## **Management Summary**

Das vorliegende Forschungsdesign bildet die konzeptionelle Grundlage für die geplante Bachelorarbeit, die den Einfluss unterschiedlicher Darstellungsweisen von KI-Konfidenzwerten auf das Vertrauen und die Technologieakzeptanz bei Large Language Model-basierten Assistenzsystemen untersuchen soll. Angesichts der rasanten Verbreitung von KI-Assistenten wie ChatGPT mit über 700 Millionen wöchentlich aktiven Nutzern adressiert die geplante Studie eine zentrale Forschungslücke: Wie beeinflusst die Art der Transparenzkommunikation das Nutzervertrauen in KI-Systeme? Diese Frage gewinnt dadurch an Relevanz, dass nahezu die Hälfte aller KI-Anfragen der Informationsbeschaffung und praktischen Anleitungen dient..

Das vorgeschlagene Forschungsvorhaben stützt sich auf das Artificial Intelligence Technology Acceptance Model (AI-TAM) von Baroni et al., welches das klassische TAM-Modell von Davis um KI-spezifische Konstrukte wie Vertrauen in KI und Kollaborationsintention erweitert. Diese theoretische Fundierung wird durch die Integration des Attribute Framing-Effekts nach Levin und Gaeth ergänzt, um zu untersuchen, wie identische Konfidenzinformationen durch unterschiedliche Valenz-Darstellung die Wahrnehmung und das Vertrauen beeinflussen. Methodisch sieht das Design ein experimentelles Between-Subjects-Verfahren mit drei Bedingungen vor: positives Framing der Konfidenzwerte als Zuverlässigkeit, negatives Framing als Unsicherheit sowie eine Kontrollgruppe ohne Konfidenzanzeige. Die geplante Manipulation soll während der natürlichen Interaktion mit einem digitalen Assistenten erfolgen.

Das Hypothesengerüst umfasst zehn Annahmen, die sowohl direkte Framing-Effekte als auch komplexe Mediations- und Moderationsbeziehungen zwischen den AI-TAM-Konstrukten postulieren.

Die vorgesehene Datenerhebung kombiniert mehrere methodische Zugänge in einer Triangulation. Standardisierte Online-Befragungen mit validierten Skalen sollen in drei Phasen – vor, während und nach der KI-Interaktion – die relevanten Konstrukte wie Explainable AI Trust, Behavioral Intention und die klassischen TAM-Variablen

erfassen. Parallel dazu ist eine automatisierte Verhaltensbeobachtung durch System-Logging geplant, die Nutzungsdaten wie Interaktionshäufigkeit, Session-Dauer und tatsächliche Accuracy Scores dokumentiert.

Die geplante Stichprobe soll aus Nutzern eines digitalen Assistenten rekrutiert werden, wobei eine randomisierte Zuweisung zu den Experimentalbedingungen direkt bei der ersten Interaktion erfolgt. Die Auswertung erfolgt mittels Strukturgleichungsmodellierung (SEM), welche die simultane Prüfung aller angenommenen Beziehungen zwischen den Konstrukten ermöglicht. Dabei werden sowohl direkte Effekte des Framings auf das Vertrauen als auch indirekte, möglicherweise medierte Effekte auf die Nutzungsintention analysiert, unter Kontrolle demografischer und Störvariablen.

Die potenzielle Relevanz dieser Forschung liegt in der praktischen Anwendbarkeit für die Gestaltung transparenter KI-Applikationen. Während KI-Anbieter rechtlich verpflichtet sind, auf mögliche Fehler hinzuweisen, gibt es bislang wenig evidenzbasiertes Wissen darüber, wie diese Kommunikation optimal gestaltet werden sollte.

## Inhaltsverzeichnis

<b>Relevanz-Einordnung</b>	<b>1</b>
<b>Theoretische Einbettung</b>	<b>2</b>
Technology Acceptance Model . . . . .	2
Artificial Intelligence-Technology Acceptance Model . . . . .	3
Framing-Effekt . . . . .	6
Forschungsfrage . . . . .	7
Latente Konstrukte . . . . .	7
Hypothesenübersicht . . . . .	9
Ausformulierte Hypothesen . . . . .	9
<b>Forschungsdesign</b>	<b>11</b>
Experimentelles Design . . . . .	12
Between-Subject Design . . . . .	12
Ablauf Experiment . . . . .	12
Pre-Interaktionsphase «digitaler Assistent» . . . . .	13
Interaktionsphase «digitaler Assistent» . . . . .	13
Post-Interaktionsphase «digitaler Assistent» . . . . .	13
Stimulus-Konzept . . . . .	14
Konkrete Stimulus-Beispiele . . . . .	15
Manipulationscheck Stimulus . . . . .	15
Methodentriangulation . . . . .	16
Online-Befragung als Rahmenstruktur . . . . .	16
Verhaltensbeobachtung durch System-Logging . . . . .	17
Experimentelle Manipulation im Feldkontext . . . . .	17
Methodenintegration . . . . .	17
Abgrenzung des Forschungsdesigns . . . . .	18
Inhaltlich . . . . .	18
Methodisch . . . . .	18

Operationalisierung Konstrukte . . . . .	19
Stichprobe/Feldzugang . . . . .	19
Beschreibung Stichprobe . . . . .	19
Rekrutierung Stichprobe . . . . .	19
Rekrutierung Stichprobe Pre-Test . . . . .	20
Pre-Test: Durchführung und Analyse . . . . .	20
Kritische Fragen und Überlegungen . . . . .	20
Stimulus-Design . . . . .	20
Experiment Ablauf . . . . .	20
Pre-Test . . . . .	20
<b>Ausgangslage</b>	<b>21</b>
<b>Machbarkeit</b>	<b>21</b>
<b>Zielsetzung</b>	<b>22</b>
<b>Arbeitsplan</b>	<b>23</b>
<b>Commented Review</b>	<b>24</b>
<b>Reflection Work</b>	<b>24</b>
<b>Quellenverzeichnis</b>	<b>25</b>
<b>Analyse</b>	<b>27</b>
Pretest-Analyse . . . . .	27
Reliabilitätsanalyse (Cronbach's Alpha) . . . . .	27
Manipulationscheck (t-test unabh. Stichproben) . . . . .	27
Hauptanalyse: Deskriptive Statistik . . . . .	27
Stichprobenbeschreibung und Gruppenvergleichbarkeit . . . . .	27
Skaleneigenschaften und Verteilungsanalysen . . . . .	28
Manipulationscheck . . . . .	28

Nutzungsverhalten und Systemvariablen . . . . .	28
Hauptanalyse: Strukturgleichungsmodellierung (SEM) . . . . .	29
Parameterschätzung und Modell-Fit . . . . .	29
Evaluation des Messmodells . . . . .	30
Strukturmodell und Hypothesentests . . . . .	30
Indirekte Effekte und Mediation (H9) . . . . .	30
Kontrolle für Störvariablen . . . . .	31

## Relevanz-Einordnung

Mit der Veröffentlichung von ChatGPT von OpenAI im Jahr 2022 (Cunningham et al., 2025) wurde eine technologische Wende eingeleitet. Bereits heute vereinfachen und verändern LLM-basierte Applikationen wie ChatGPT von OpenAI, Claude von Anthropic und Gemini von Google viele Tätigkeiten des (Arbeits-)Lebens. Die rasante Verbreitung dieser Technologie zeigt sich eindrücklich: Innerhalb von nur sieben Monaten im Jahr 2025 konnte OpenAI seine Nutzerbasis von 350 auf über 700 Millionen wöchentlich aktive Nutzer steigern (Cunningham et al., 2025).

Chatbots spielen im Alltag inzwischen in mehrfacher Hinsicht eine wichtige Rolle: Sie unterstützen bei der Informationsbeschaffung, geben praktische Anleitungen und bieten zum Beispiel Hilfe in der Programmierung sowie in Kreativprozessen. Dabei treten sie als tägliche Begleiter des Menschen auf: Ob durch eine bewusst durchgeführte Interaktion oder als ein im Hintergrund stattfindender, unbewusster Berührungspunkt (Cunningham et al., 2025).

Mit ihrer relativ jungen (und öffentlichkeitswirksamen) Geschichte ist die generative künstliche Intelligenz, wie viele übergreifende technologischen Veränderungen, einem technologischen- und gesellschaftlichen Adoptionsprozess ausgesetzt. Einen theoretischen Erklärungsansatz dieses Adoptionsprozesses liefert Fred Davis 1987 mit seinem Werk «User acceptance of information systems: the technology acceptance model (TAM)». In seiner Arbeit legt Davis den Fokus auf die wahrgenommene Nützlichkeit («Perceived Usefulness») und die Einfachheit der Nutzung («Ease of Use»), woraus die Verhaltensintention («Behavioural Intention») abgeleitet wird (Davis, 1989). Im Kontext von generativer KI, oder künstlicher Intelligenz im Allgemeinen, ist jedoch der Aspekt des Vertrauens in die Technologie von besonderer Bedeutung. Neben Nützlichkeit und Einfachheit stellt die Vertrauensfrage den Aspekt dar, ob künstlicher Intelligenz vertrauenswürdig ist. Sämtliche grossen Anbieter wie ChatGPT, Claude von Anthropic, etc. weisen vor- sowie während der Nutzung ausdrücklich darauf hin, dass ihre «Foundational Models» und konsequenterweise KI-Assistenten die auf diesen Modellen basieren, fehlerhaft sein

können. Diese Fehleranfälligkeit sowie zusätzliche Vorbehalte, wie die Angst vor Jobverlust, Bedenken hinsichtlich der Privatsphäre oder ethische Fragen (Li & Huang, 2020), erfordern die Integration und Erfassung von «Vertrauen» als eigenständiges Konstrukt in möglichen theoretischen Modellen.

Das bestehende TAM-Modell von Davis (1989) kann durch verschiedene Faktoren wie Vertrauen in KI (XAIT, «Explainable AI Trust»), Intention zur Kollaboration mit künstlicher Intelligenz («Collaborative Intention») erweitert werden. Baroni et al. (2022) schufen mit ihrer Erweiterung des TAM-Modells «AI-TAM» in ihrer Studie einen neuen Erklärungsbeitrag zum Thema «Vertrauen in KI». Die theoretische Einbettung sowie das Forschungsdesign orientieren sich an diesem erweiterten AI-TAM-Modell. In den folgenden Schritten wird die genaue Integration des AI-TAM Modells in das geplante Experiment beschrieben.

### **Theoretische Einbettung**

Modelltheoretisch knüpft die vorliegende Arbeit an frühere Studien in den Bereichen Vertrauen in künstliche Intelligenz/technologische Veränderungen, wahrgenommene Nützlichkeit sowie Benutzerfreundlichkeit und die daraus abgeleitete Nutzungsabsicht an. Als theoretische Grundlage dient zunächst das Technology Acceptance Model (TAM) von Fred Davis aus dem Jahr 1987, welches die Rahmenbedingungen zur Analyse von Adoptionsprozessen neuer Technologien schafft (Davis, 1989). Den zweiten Baustein liefert die Erweiterung des TAM-Modells durch Baroni et al. (2022). Diese ergänzt das Modell um zusätzliche Faktoren wie das Vertrauen in KI-gestützte Assistenten und bildet diese im Artificial Intelligence Technology Acceptance Model (AI-TAM) ab (Baroni et al., 2022). Zuletzt wird der Framing-Effekt theoretisch beleuchtet, da dieser für die gewählte Stimulus-Wahl relevant ist. Konkret wird dabei die Form des Attribute-Framing-Effekts betrachtet (Druckman, 2001; Freling et al., 2014).

#### ***Technology Acceptance Model (Davis, 1989)***

Das TAM wurde entwickelt, um die mangelnde Nutzerakzeptanz von Informationssystemen zu adressieren, die als Haupthindernis für den Erfolg neuer

Technologien identifiziert wurde. Davis untersuchte 112 Angestellte und Manager eines grossen nordamerikanischen Unternehmens, die zwei unterschiedliche Softwaresysteme nutzten - ein elektronisches Mailsystem und einen Texteditor. Das Modell basiert auf der Attitude-Paradigm aus der Psychologie, speziell auf Fishbein und Ajzens Theory of Reasoned Action (Fishbein & Ajzen, 1975). TAM besagt, dass die tatsächliche Systemnutzung durch die Verhaltensintention bestimmt wird, welche von der Einstellung zur Nutzung abhängt.

Diese Einstellung wird durch zwei zentrale Konstrukte geprägt:

- Perceived Usefulness – «the degree to which an individual believes that using a particular system would enhance his or her job performance» – und
- Perceived Ease of Use – «the degree to which an individual believes that using a particular system would be free of physical and mental effort».

Die Studie zeigte, dass Perceived Usefulness etwa 50% einflussreicher auf die Nutzung war als Ease of Use, wobei das Modell 36% der Varianz in der tatsächlichen Nutzung erklären konnte.

Während TAM erfolgreich die Akzeptanz traditioneller Informationssysteme erklärt, erweist es sich für KI-basierte Systeme als unzureichend. KI-Systeme unterscheiden sich durch ihre probabilistische Natur und inhärente Unsicherheit - Eigenschaften, die Vertrauen zu einem Faktor machen, der im ursprünglichen TAM nicht berücksichtigt wird. Zudem werden KI-Systeme nicht nur als Werkzeuge, sondern oft auch als kollaborative Partner wahrgenommen, was neue Dimensionen der Mensch-Maschine-Interaktion eröffnet. Diese Lücke adressieren Baroni et al. (2022) mit ihrer Erweiterung des TAM-Modells (Baroni et al., 2022).

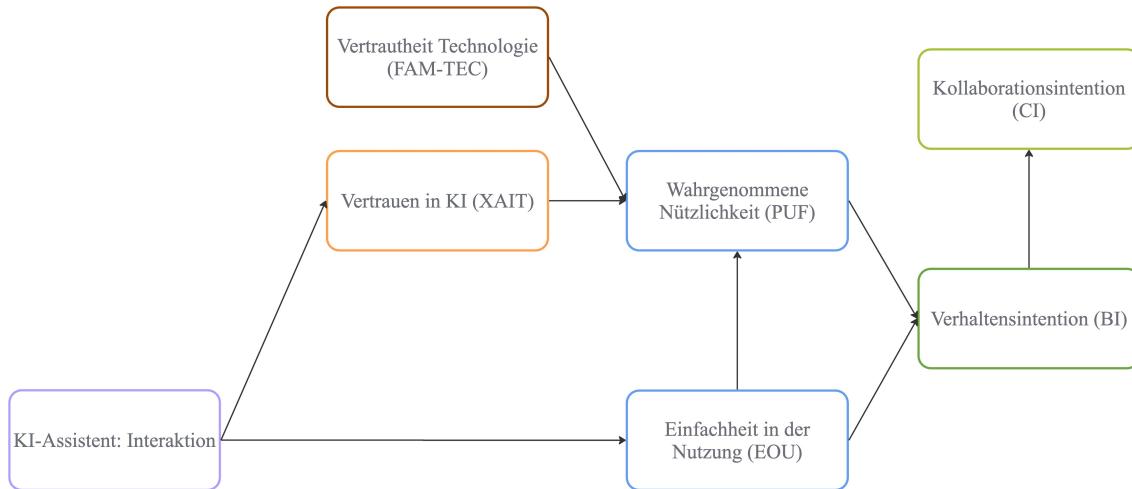
### ***Artificial Intelligence-Technology Acceptance Model (Baroni et al., 2022)***

**Notwendigkeit der TAM-Erweiterung für KI-Systeme.** Das von Davis (1989) entwickelte Technology Acceptance Model (TAM, (Davis, 1989)) basiert auf der Theory of Reasoned Action und erklärt Technologieakzeptanz durch die Faktoren der wahrgenommenen Nützlichkeit (Perceived Usefulness) und der wahrgenommenen

Benutzerfreundlichkeit (Perceived Ease of Use). Während das TAM die Adoption traditioneller Informationssysteme bereits untersucht hat, erweist es sich für KI-basierte Systeme als unzureichend. Der Grund dafür liegt in den Unterschieden von KI-Systemen: Ihre probabilistische Natur, die Unsicherheit und ihre Wahrnehmung als kollaborative Partner anstelle reiner Werkzeuge. Diese Eigenschaften machen Vertrauen zu einem Faktor, der im ursprünglichen TAM nicht abgebildet wird. Darüber hinaus erfordern «Human-in-the-Loop»-Ansätze, dass Nutzer aktiv zur Verbesserung der KI beitragen – eine Dimension der Kollaboration, die das klassische Modell ebenfalls nicht berücksichtigt.

**Erweiterungen des TAM zum AI-TAM.** Baroni et al. (2022) erweiterten das TAM um drei zusätzliche Konstrukte: «Explainable AI Trust» (Vertrauen in KI) aus der Literatur zu «Explainable AI» (XAI), «Collaborative Intention» (Kollaborationsabsicht) zur Messung der Bereitschaft zur Teilnahme an «Human-in-the-Loop»-Mechanismen sowie die Vertrautheit mit der Technologie und dem Anwendungskontext. Vertrauenskonstrukt basiert auf der Arbeit von Hoffman et al. (Hoffman et al., 2019). Dabei misst «Explainable AI Trust» die Zuversicht in die Ergebnisse der KI. Die «Collaborative Intention» erfassst die Bereitschaft der Nutzer, aktiv zur Verbesserung der KI beizutragen, was ein kritischer Faktor für «Human-in-the-Loop»-Systeme ist. Dieses übt einen Einfluss auf Kernkonstrukte des TAM aus.

**Validierung durch die BumpOut-Studie (Baroni et al., 2022).** Das AI-TAM wurde mit der Anwendung «BumpOut» validiert, einer KI-gestützten App zur Schadensmeldung bei Autounfällen. Die Studie umfasste 400 Teilnehmende in zwei Crowdsourcing-Kampagnen unter unterschiedlichen experimentellen Bedingungen: einer fehlerfreien KI versus einer teilweise fehlerhaften KI. Die App analysiert dabei Schadensbilder automatisch, wobei die Nutzer die von der KI getroffenen Identifikationen bestätigen oder korrigieren können. Die Ergebnisse zeigten hohe Werte für die wahrgenommene Nützlichkeit und Benutzerfreundlichkeit, während die Funktionsfähigkeit der KI nur einen minimalen Einfluss hatte. Besonders bedeutsam



**Abbildung 1**

*Erweitertes TAM-Modell: Artificial Intelligence-Technology Acceptance Model (Baroni et al., 2022)*

war die starke Korrelation zwischen der Nutzungsabsicht («Behavioral Intention») und der Kollaborationsabsicht («Collaborative Intention»). Dies bestätigt, dass Nutzer, die bereit sind, die App zu verwenden, auch bereit sind, zur Verbesserung der KI beizutragen. Das AI-TAM eignet sich daher auch für die Untersuchung der Akzeptanz von Large Language Models (LLMs), da diese Systeme die gleichen kritischen Charakteristika aufweisen: probabilistische Ausgaben, inhärente Unsicherheit und die Notwendigkeit von Nutzervertrauen. LLMs werden zunehmend als kollaborative Partner wahrgenommen, deren Ergebnisse oft Nutzerfeedback erfordern. Insbesondere die XAI-Konstrukte sind hier relevant, da Nutzer nachvollziehen müssen, warum ein LLM eine bestimmte Antwort generiert.

Die Verbindung des AI-TAM mit dem Konzept des Attribute Framings eröffnet neue Forschungsperspektiven. Framing-Effekte könnten insbesondere das

XAIT-Konstrukt beeinflussen. So dürfte die Präsentation von KI-Fähigkeiten als Gewinn («95 % Genauigkeit») im Vergleich zu einer Darstellung als Unsicherheit («5 % Fehlerrate») das Vertrauen in die KI («Explainable AI Trust») direkt verändern. Gemäss dem AI-TAM-Modell beeinflusst dieser Faktor wiederum die Nutzungsabsicht («Behavioral Intention»). Für Experimente mit Large Language Models (LLMs) bedeutet dies, dass die Art der Leistungsdarstellung, beispielsweise durch das Attribute Framing der Modellfähigkeiten, die Nutzerakzeptanz beeinflussen könnte. Das AI-TAM bietet hierbei den methodischen Rahmen, um diese Effekte auf den relevanten Dimensionen präzise zu messen.

### ***Framing-Effekt***

Der Framing-Effekt, erstmals von Kahneman und Tversky in ihrer Prospect Theory beschrieben, zeigt, dass Entscheidungen davon beeinflusst werden, wie Informationen präsentiert werden (Tversky & Kahneman, 1986). Der Framing-Effekt zeigt unter anderem, wie identische Szenarien zu unterschiedlichen Präferenzen führen, je nachdem ob sie in Gewinn- oder Verlustbegriffen formuliert werden. Während sich die frühe Forschung auf riskante Entscheidungen konzentrierte, erweiterte sich das Konzept auf verschiedene Framing-Typen wie Risky Choice Framing, Goal Framing und Attribute Framing.

Zusätzlich untersuchte Freling et al. (2014) in ihrer Meta-Analyse 107 Studien zum Attribute Framing und entwickelten dabei eine theoretische Integration mittels Construal Level Theory (CLT). Ihre zentrale Erkenntnis: Die Effektivität von Attribute Framing hängt von der Kongruenz zwischen dem Abstraktionsniveau (Construal Level) des Frames und der psychologischen Distanz des Bewertenden zum geframten Event ab (Freling et al., 2014).

**Attribute Framing nach Levin & Gaeth (1988) und Dolgopolova et al. (2022).** Attribute Framing unterscheidet sich von anderen Framing-Typen, da hier ein einzelnes Attribut in äquivalenten aber unterschiedlich valenten Begriffen beschrieben wird. Levin und Gaeth demonstrierten dies mit Hackfleisch, das entweder als «75% mager» oder «25% fett» beschrieben wurde (Levin & Gaeth, 1988). Der Attribute

Framing-Effekt manifestiert sich in einer valenz-konsistenten Verschiebung: Positive Frames führen zu günstigeren Bewertungen als negative. Ihre Studie zeigte zudem, dass direkte Produkterfahrung den Framing-Effekt abschwächt - ein Befund, der durch ein Averaging-Modell erklärt wird, bei dem zusätzliche Informationsquellen den relativen Einfluss einzelner Frames reduzieren.

Dolgopolova et al. (2022) spezifisch auf Lebensmittelentscheidungen und fanden Effekte für Einstellungen versus Intentionen. Während Gain-Frames signifikant positivere Einstellungen erzeugten, war der Effekt auf Kaufintentionen nahe null und nicht signifikant. Mehrere Moderatoren wurden identifiziert: Gain-Frames, Interaktionsterme, spezifische Produkte und Studentenstichproben beeinflussten signifikant die Ergebnisse. Diese Befunde unterstreichen die Komplexität des Attribute Framing bei Lebensmitteln, wo zeitliche Diskontierung und die Verzögerung zwischen Konsum und Gesundheitskonsequenzen eine Rolle spielen (Dolgopolova et al., 2022).

Der Attribute Framing-Effekt ist für die Untersuchung der KI-Akzeptanz relevant, da KI-Systeme durch ihre Fähigkeiten (Gain-Frame: «95% Genauigkeit») oder Limitationen (Loss-Frame: «5% Fehlerrate») charakterisiert werden können. Im Kontext des erweiterten TAM-Modells (Davis, 1989) könnte Attribute Framing die Wahrnehmung von Perceived Usefulness und Vertrauen in KI beeinflussen. Die Präsentation von KI-Funktionen als Gewinne («erhöht Produktivität um 30%») versus Verluste («30% manuelle Arbeit bleibt erforderlich») könnte unterschiedliche Akzeptanzmuster erzeugen.

### **Forschungsfrage**

Wie beeinflusst die Framingdarstellung von KI-Konfidenzwerten (positiv vs. negativ) das Vertrauen in KI-generierte Antworten und die daraus resultierende Technologieakzeptanz in LLM-basierten Assistenzsystemen?

### **Latente Konstrukte**

Die latenten Konstrukte werden mittels einer Online-Befragung vor, während und nach der Nutzung der KI-Assistenz erhoben. Die verwendeten Konstrukte stammen grossteils aus dem Technology Acceptance Model (TAM) von Davis sowie aus der

Erweiterung dieses Modells durch Baroni et al. (2022). Diese Erweiterung ergänzt das bestehende TAM um KI-relevante Faktoren wie das Vertrauen in erklärbare KI («Explainable AI Trust», XAIT), die Kollaborationsabsicht («Collaborative Intention, CI») und die technologische Vorerfahrung («Familiarity with Technology», FAM-TEC») (Baroni et al., 2022).

**Tabelle 1**

*Identifizierte latente Konstrukte zum Einsatz von AI-TAM*

Abkürzung	Name	Definition	Quelle
XAIT	Explainable AI trust	Vertrauen in die generierte Antwort und die LLM-Lösung	(Baroni et al., 2022)
BI	Behaviorial Intention	Verhaltensintention die LLM-Lösung zu Nutzen	(Baroni et al., 2022; Davis, 1989)
CI	Collaborative Intenti- on	Kollaborationsintention zur digitalen Assis- tenz	(Baroni et al., 2022; Davis, 1989)
PUF	Perceived Usefulness	Wahrgenommene Nützlichkeit der ge- nerierten Antwort	(Baroni et al., 2022; Davis, 1989)
EOU	Ease of Use	Einfachheit in der Nutzung der KI- Applikation	(Baroni et al., 2022; Davis, 1989)
FAM-TEC	Familiarity with Tech- nology	Vertrautheit in der Nutzung von KI- Technologie	(Baroni et al., 2022; Davis, 1989)

**Tabelle 2***Aufgelistete Hypothesen im Rahmen der Bachelor-Arbeit-Vorstudie*

Hypothese	Pfad	Richtung	Theorie
H1a	Dummy_Pos (Stimulus) → XAIT	+	Attribute Frame
H1b	Dummy_Neg (Stimulus) → XAIT	-	Attribute Frame
H2	XAIT → PUF	+	AI-TAM-Modell
H3	XAIT → EOU	+	AI-TAM-Modell
H4	PUF → BI	+	AI-TAM-Modell
H5	EOU → BI	+	AI-TAM-Modell
H6	EOU → PUF	+	AI-TAM-Modell
H7	BI → CI	+	AI-TAM-Modell
H8	FAMTEC → PUF	+	AI-TAM-Modell

**Kovariate**

ACTS	ACTS → XAIT	kontrolliert	-
------	-------------	--------------	---

**Mediation**

H9	Framing → XAIT → (PUF, EOU) → BI	Indirekt	-
----	----------------------------------	----------	---

Anmerkung: Alle Pfade werden simultan im Strukturgleichungsmodell (SEM) geschätzt

**Hypothesenübersicht*****Ausformulierte Hypothesen*****Haupthypothesen (Framing-Effekte).**

- H1a: Die positive Darstellung des Accuracy Scores (z.B. «Diese Antwort ist zu 80% korrekt») führt zu einem höheren AI Output Trust als die Kontrollbedingung ohne Score-Anzeige.
- H1b: Die negative Darstellung des Accuracy Scores (z.B. «Diese Antwort hat eine 20% Fehlerwahrscheinlichkeit») führt zu einem niedrigeren AI Output Trust als die Kontrollbedingung ohne Score-Anzeige.

### **AI-TAM Kernbeziehungen.**

- H2: AI Output Trust hat einen positiven Einfluss auf die Perceived Usefulness.  
Nutzer, die den AI-Ausgaben vertrauen, bewerten das System als nützlicher für ihre Aufgaben.
- H3: AI Output Trust hat einen positiven Einfluss auf die Perceived Ease of Use.  
Vertrauen in die AI-Ausgaben reduziert die wahrgenommene kognitive Belastung bei der Systemnutzung.

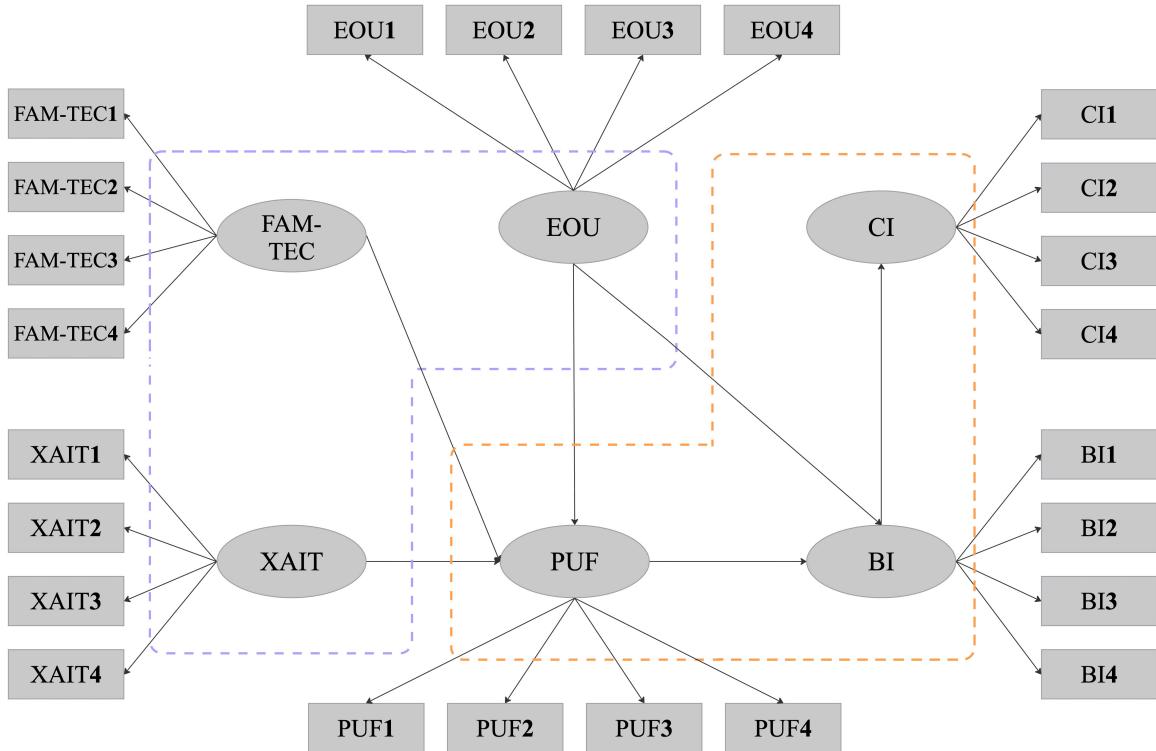
### **TAM-Standardbeziehungen.**

- H4: Die Perceived Usefulness hat einen positiven Einfluss auf die Behavioral Intention. Je nützlicher Nutzer Alva einschätzen, desto höher ist ihre Absicht, das System zukünftig zu nutzen.
- H5: Die Perceived Ease of Use hat einen positiven Einfluss auf die Behavioral Intention. Eine als einfach wahrgenommene Nutzung erhöht die Intention zur zukünftigen Systemnutzung.
- H6: Die Perceived Ease of Use hat einen positiven Einfluss auf die Perceived Usefulness. Systeme, die einfach zu nutzen sind, werden als nützlicher wahrgenommen.
- H7: Die Behavioral Intention hat einen positiven Einfluss auf die Collaborative Intention. Nutzer, die beabsichtigen Alva zu nutzen, zeigen auch eine höhere Bereitschaft zur kollaborativen Zusammenarbeit mit dem AI-System.
- H8: Familiarity with Technology hat einen positiven Einfluss auf die Perceived Usefulness. Nutzer, die mit KI-Technologie vertraut sind, schätzen die Nützlichkeit der digitalen Assistenz höher ein.

### **Mediation.**

- H9: Der Effekt des Framings auf die Behavioral Intention wird durch Explainable AI Trust partiell oder vollständig mediert. Bei niedrigen Explainable AI

Trust-Werten ist der Unterschied zwischen positivem und negativem Framing grösser als bei hohen Explainable AI Trust-Werten.



**Abbildung 2**

*Strukturmodell für AI-TAM, Teil des Strukturgleichungsmodells, inkl. Hypothesen, inkl. Stimulus*

### Forschungsdesign

Das Experiment untersucht den Einfluss von Accuracy-Framing auf die Technologieakzeptanz von KI-Systemen, basierend auf dem AI-TAM-Modell. In einem 3x2 Between-Subjects-Design wird die Darstellung von Konfidenzwerten (positiver Frame vs. negativer Frame) bei variierenden Accuracy-Scores (hoch, mittel, niedrig) manipuliert. Die experimentelle Manipulation erfolgt während der realen Interaktion mit einem KI-Assistenten.

## **Experimentelles Design**

### ***Between-Subject Design***

Das Experiment nutzt ein Between-Subjects-Design mit drei Experimentalgruppen, um den Einfluss des Accuracy-Framings auf das Vertrauen in KI-Systeme zu untersuchen.

- Unabhängige Variable: Das Accuracy-Framing mit zwei Ausprägungen
  - Positiver Frame: Gain-Darstellung (positive «Zuversicht»)
  - Negative Frame: Loss-Darstellung (negative «Zuversicht»)
- Unabhängige Variable: Das Accuracy-Framing mit drei Stufen:
  - Hoch
  - Mittel
  - Niedrig

Diese Designstruktur wird in der Literatur als Between-Subjects-Design mit Kovariate bezeichnet, wobei die kontinuierliche Variable als statistische Kontrollvariable dient (Kim & Song, 2022). Die Kombination eines kategorialen Faktors mit einer kontinuierlichen Variable ermöglicht die Untersuchung von Haupt- und Interaktionseffekten, während gleichzeitig die natürliche Varianz der KI-Performance kontrolliert wird

## **Ablauf Experiment**

Das geplante Experiment ist in drei Phasen gegliedert: eine Phase vor, eine während und eine nach der Interaktion mit dem KI-Assistenten. In jeder dieser Phasen werden die benötigten Daten erhoben. Dies geschieht entweder durch direkte Nutzerbefragung mittels Bewertungsfragen oder durch automatische Berechnung und Speicherung im Hintergrund, wie es beispielsweise beim «Accuracy Score» der Fall ist.

Da die initiale Befragung bereits stattfindet, bevor die Teilnehmenden mit dem System interagieren, wird diese bewusst kurz gehalten. Ziel ist es, die Abbruchrate zu minimieren und eine hohe Abschlussrate des gesamten Experiments zu fördern.

**Tabelle 3***Experiment-Design, vorhandene Experimentalbedingungen*

<b>Bedingung</b>	<b>Gruppe</b>	<b>Manipulation</b>	<b>Beispiel</b>
Positive-Frame	Gruppe 1	Score wird als Konfidenz/Zuverlässigkeit dargestellt	«Antwortsicherheit: 80%» oder «Antwortsicherheit zu 80% zuverlässig»
Negative-Frame	Gruppe 2	Score wird als Unsicherheit/Fehlerwahrscheinlichkeit oder dargestellt	Antwortunsicherheit: 20% oder «Diese Antwort hat eine Fehlerwahrscheinlichkeit von 20%»
Kontrollgruppe	Gruppe 3	Kein Score wird angezeigt (Status Quo)	-

***Pre-Interaktionsphase «digitaler Assistent»***

- Einführung & Einwilligung
- Erhebung «FAM-TEC»-Konstrukt

***Interaktionsphase «digitaler Assistent»***

- Anzahl Interaktionen
- Tatsächlicher Accuracy-Score von Antworten
- Dauer der Sitzungen
- Anonymisierte, qualitative Inhalte wie Prompts und Antworten
- Bewertung («Rating»)
- LLM-Einschätzung («Faithfulness Score»)

***Post-Interaktionsphase «digitaler Assistent»***

- Erhebung «AI-TAM-Konstrukte»

**Tabelle 4***Experimentelles 3x2 Design: Manipulation von Framing und Accuracy*

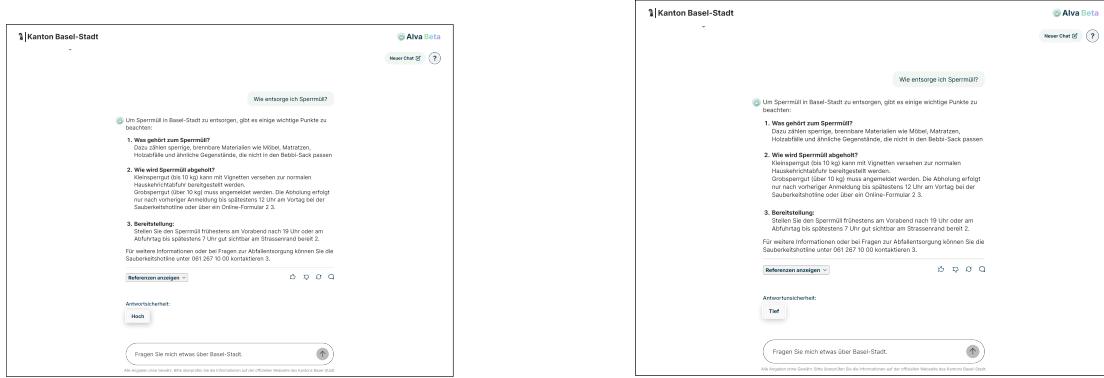
Framing (Unabhängige Variable 1)		
Accuracy (UV 2)	Positiver Frame	Negativer Frame
<b>Hoch</b>	Sicherheit: Hoch	Unsicherheit: Tief
<b>Mittel</b>	Sicherheit: Mittel	Unsicherheit: Mittel
<b>Niedrig</b>	Sicherheit: Tief	Unsicherheit: Hoch

Anmerkung. Die Kontrollgruppe (kein Score) ist in diesem 3x2 Design nicht abgebildet.

- Erhebung «XAIT»-Konstrukt
- Erhebung «TAM-Konstrukte»
- Erhebung «PUF»-Konstrukt, Teil 2
- Erhebung «EOU»-Konstrukt
- Erhebung «BI»-Konstrukt
- Erhebung «CI»-Konstrukt
- Erhebung soziodemografischer Daten (Kontrollvariablen)
  - Alter
  - Geschlecht
  - Höchster Bildungsabschluss

### Stimulus-Konzept

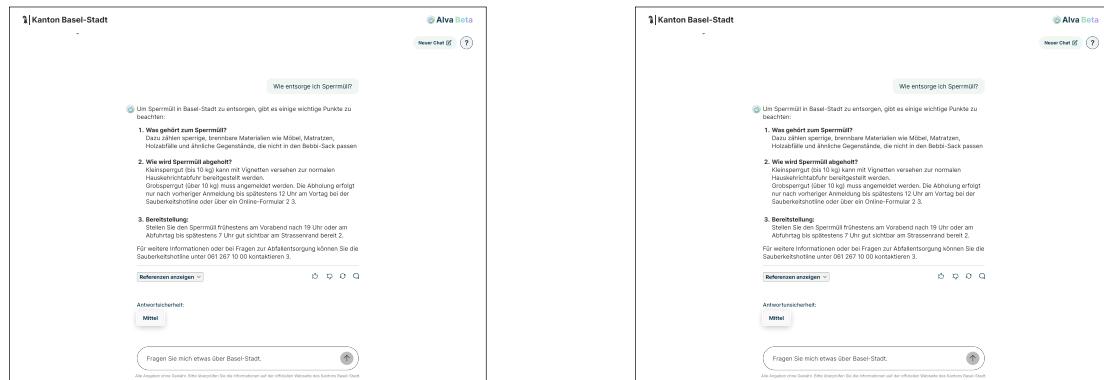
Der Stimulus besteht aus der visuellen und textlichen Darstellung des AI-Accuracy Scores direkt nach jeder LLM-Antwort. Die Manipulation erfolgt in Echtzeit während der natürlichen Interaktion mit dem digitalen Assistenten.



(a) Stimulus: Sicherheit hoch

**Abbildung 3***Stimulus-Design: Hohe Sicherheit vs. Tiefe Unsicherheit*

(b) Stimulus: Unsicherheit tief



(a) Stimulus: Sicherheit mittel

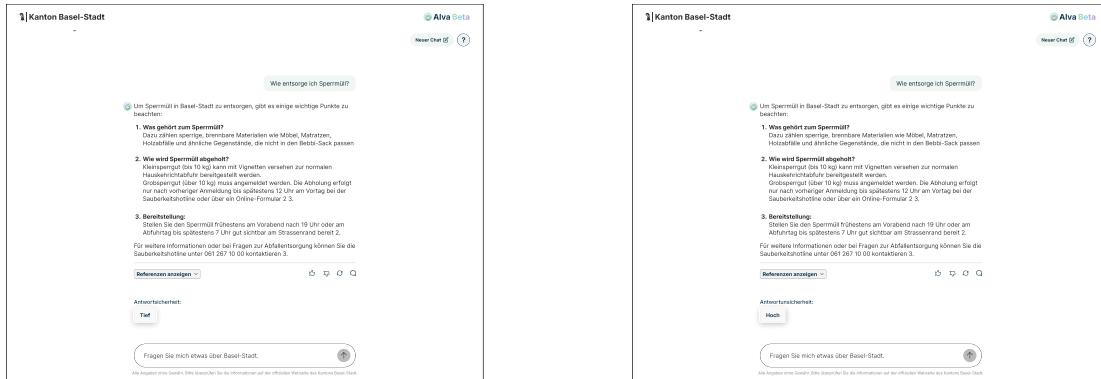
**Abbildung 4***Stimulus-Design: Mittlere Sicherheit vs. Mittlere Unsicherheit*

(b) Stimulus: Unsicherheit mittel

## Konkrete Stimulus-Beispiele

### Manipulationscheck Stimulus

Als Manipulationscheck werden die Probanden post-experimentell gefragt, ob und in welcher Form ihnen Informationen zur Zuverlässigkeit der Antworten angezeigt wurden, um sicherzustellen, dass die experimentelle Manipulation wahrgenommen wurde.

(a) *Stimulus: Sicherheit tief*(b) *Negativer Frame / Niedrige Sicherheit***Abbildung 5***Übersicht der Stimulus-Designs für das 3x2 Experiment***Methodentriangulation**

Das vorliegende Forschungsdesign kombiniert verschiedene Methoden in einer Methodentriangulation, um die Forschungsfrage nach dem Einfluss von KI-Transparenz auf Vertrauen zu beantworten.

Die Triangulation erfolgt auf drei Ebenen:

- experimentelle Manipulation des Accuracy Framings als Between-Subject-Design gewährleistet die interne Validität durch randomisierte Zuweisung.
- Die natürliche Beobachtung während der tatsächlichen LLM-Nutzung erhöht die ökologische Validität, da Nutzer eigene Fragen in realistischen Anwendungskontexten stellen.
- Die standardisierte Befragung mittels validierter Skalen aus dem AI-TAM-Modell ermöglicht die reliable Messung latenter Konstrukte.

***Online-Befragung als Rahmenstruktur***

Die gesamte Datenerhebung erfolgt über eine webbasierte Plattform, die Pre-Interaction-Messungen (demografische Daten, AI-Vorerfahrung), Post-Interaction-Messungen (AI-TAM-Konstrukte) und die experimentelle Randomisierung steuert. Die Verwendung etablierter Skalen aus dem TAM (Davis,

1989) und AI-TAM (Baroni et al., 2022) gewährleistet die Vergleichbarkeit mit bestehender Forschung. Die standardisierten Items werden auf 5-stufigen Likert-Skalen gemessen.

### ***Verhaltensbeobachtung durch System-Logging***

Während der Interaktion mit dem digitalen Assistenten werden automatisiert Verhaltensdaten erfasst: Anzahl der Interaktionen, Session-Dauer(, Fragentypen). Diese Messung liefert weitere Verhaltensindikatoren. Der tatsächliche Accuracy Score wird vom LLM-System für jede Antwort berechnet und protokolliert, wodurch eine kontinuierliche, objektive Performanz-Metrik entsteht, die als Kovariate in die Analysen eingeht.

### ***Experimentelle Manipulation im Feldkontext***

Die Framing-Manipulation wird während der natürlichen Nutzung implementiert. Diese Einbettung des Experiments in den realen Anwendungskontext entspricht einem natürlichen Experiment das externe bei ausreichender interner Validität bietet.

### ***Methodenintegration***

Die Integration der verschiedenen Datenquellen erfolgt auf Analyseebene: Die experimentelle Gruppenzugehörigkeit (Framing) wird mit den Befragungsdaten (Trust, TAM-Konstrukte) und den Systemdaten (Accuracy Score) in einem gemeinsamen Datensatz zusammengeführt. Diese Triangulation ermöglicht:

- Konvergenz-Validierung: Trust-Ratings während der Interaktion (Single-Item) werden mit Post-Interaction Trust-Skalen (Multi-Item) korreliert
- Komplementäre Erkenntnisse: Subjektive Wahrnehmungen (Befragung) werden mit objektiven Metriken (System-Logs) kontrastiert
- Moderation/Mediation: Die Interaktion zwischen experimenteller Manipulation und natürlicher Variation kann analysiert werden

## **Abgrenzung des Forschungsdesigns**

Die vorliegende Studie fokussiert auf die valenzorientierte Darstellung von KI-Leistungsmetriken (Attribute Framing) und deren Einfluss auf Vertrauen und Technologie-akzeptanz im Kontext des AI-TAM-Modells.

### ***Inhaltlich***

- Andere Framing-Typen: Die Studie beschränkt sich auf Attribute Framing und untersucht nicht Risky Choice Framing oder Goal Framing.
- Langzeiteffekte (keine Längsschnittstudie): Gemessen wird die Nutzungsabsicht (Behavioral Intention), nicht die tatsächliche Systemnutzung über längere Zeiträume. Der Intention-Behavior-Gap wird nicht untersucht.
- Alternative Transparenzmechanismen: Neben der Score-Darstellung existieren weitere Transparenzmöglichkeiten (Erklärungen, Quellenangaben, Visualisierungen), die nicht Gegenstand dieser Arbeit sind.
- Kontextübergreifende Generalisierung: Die Untersuchung findet im Verwaltungskontext statt. Ob die Ergebnisse auf medizinische, kreative oder andere Anwendungsbereiche übertragbar sind, bleibt offen.

### ***Methodisch***

- Between-Subjects-Design: Jede Person erfährt nur eine Framing-Bedingung. Intraindividuelle Vergleiche sind nicht möglich.
- Quantitative Fokussierung: Die Studie nutzt standardisierte Skalen, verzichtet jedoch auf qualitative Vertiefungen wie Interviews zur Exploration der zugrunde liegenden kognitiven Prozesse.
- Natürliche Variation des Accuracy Scores: Der tatsächliche Score wird nicht experimentell manipuliert, sondern variiert basierend auf den Nutzeranfragen. Er dient als Kovariate, nicht als unabhängige Variable.

## **Operationalisierung Konstrukte**

Die Operationalisierung der Konstrukte erfolgt in Form von Bewertungsfragen mit einer 5-Punkt Likert-Skalenbewertung. Die Befragung wird mittels Onlinebefragung vor- während- und nach der Verwendung der LLM-Lösung durchgeführt. Die Likert-Skalen sind so skaliert, dass 1 jeweils die negativste Bewertung des jeweiligen Items darstellt und 5 die positivste Bewertung. Eine Item-Batterie beinhaltet zwischen ein bis sechs Items, welche das gewünschte Konstrukt erfassen sollen. Einzelne Item-Batterien beinhalten negativformulierte Items als Kontrollfragen.

Das Operationalisierungsverfahren ist theoriegeleitet, da ein Grossteil der bestehenden Items aus vorherigen Studien (Baroni et al., 2022; Davis, 1987) teilweise übernommen werden kann. Die Items aus den verschiedenen Item-Batterien (latente Konstrukte) müssen jedoch für den geplanten Anwendungsfall überarbeitet und übersetzt werden. Dies betrifft sämtliche definierten Konstrukte des definierten AI-TAM-Modells (Baroni et al., 2022).

## **Stichprobe/Feldzugang**

### ***Beschreibung Stichprobe***

Die Stichprobe ist als Gelegenheitsstichprobe zu bezeichnen, da nur potenziell Benutzer mit dem digitalen Assistenten interagieren, welche bereits wissen, dass es diesen gibt. Einschlusskriterien für die zu erhebende Stichprobe sind wie folgt ausgelegt.

### ***Rekrutierung Stichprobe***

Die Rekrutierung der Proband\*innen geschieht direkt auf der Plattform, wo der digitale Assistent integriert ist. Die Rekrutierung der Benutzer\*innen findet somit ausschliesslich digital statt. Wenn sich Benutzer\*innen entscheiden mit der digitalen Assistenz auf der Plattform zu interagieren, werden Benutzer\*innen nach Akzeptieren der Bestimmungen einer Experimentalbedingung zugewiesen.

## ***Rekrutierung Stichprobe Pre-Test***

### ***Pre-Test: Durchführung und Analyse***

#### **Kritische Fragen und Überlegungen**

##### ***Stimulus-Design***

- Stimuli unterscheiden sich in drei Dimensionen gleichzeitig: Text (Konfidenz/Unsicherheit), Farbe (grün/orange-rot) UND Icon (✓/☒)
- Manipulationscheck ausreichend bei prominenter Platzierung im Assistenten-Interface?

##### ***Experiment Ablauf***

- Es gibt mehrere Möglichkeiten die Proband\*innen in das Experiment zu holen
  - Footer
  - Button
  - Suchresultate
- Soll das Onboarding zum Experiment vor der ersten LLM-Interaktion geschehen oder nachher?
  - Sprich, ich (Proband\*in) kann zuerst eine Frage an den digitalen Assistenten richten, danach Onboarding zum Experiment
  - JA/NEIN

##### ***Pre-Test***

- Framing muss auch irgendwie gepretestet werden
- Wie sieht die Experimentalbedingung aus im Pre-Test
  - Hoch / Mittel / Tief
  - Verschiedene Sicherheitslevel
  - Alles unter 50% ist tief

## Ausgangslage

Mit der Lancierung von Alva erweitert der Kanton Basel-Stadt sein bestehendes Informationsangebot um eine KI-gestützte Interaktionsform. Bei Alva handelt es sich um einen LLM-basierten Chatbot, der auf der Technologie von ChatGPT basiert und die Inhalte der Kantonswebsite bs.ch als Wissensbasis nutzt, um Fragen der Bevölkerung in natürlicher Konversationsform zu beantworten.

Die Einführung von Alva markiert für den Kanton Basel-Stadt einen bedeutsamen Schritt: Es handelt sich um eine der ersten KI-gestützten Lösungen dieser Art im kantonalen Kontext. Die gewonnenen Erkenntnisse aus diesem Pilotprojekt sollen als Grundlage für weitere KI-basierte Initiativen des Kantons dienen.

Aktuell verzeichnet Alva täglich rund 700 aktive Nutzer, die im Durchschnitt 1.4 Interaktionen mit dem digitalen Assistenten durchführen. Nebst der erwarteten Effizienzsteigerung bei der Informationsbeschaffung ist es von besonderem Interesse zu untersuchen, inwiefern das Vertrauen in die KI-Lösung die Nutzungsabsicht beeinflusst.

## Machbarkeit

Das vorliegende Forschungsdesign basiert auf einem experimentellen Testen verschiedener Konfidenz-Werte innerhalb eines chatbot-basierten Umfeldes. Das Experiment soll im Idealfall in einer tatsächlichen Chatbot-Interaktion stattfinden, anstelle einer fragebogenbasierten Stimulus-Darbietung.

Die von Basel-Stadt entwickelte Lösung «Alva» (Kanton Basel-Stadt, 2025a) dient als zentraler digitaler Assistent bei der Bedienung und Navigation der Website des Kantons Basel-Stadt (Kanton Basel-Stadt, 2025b). Alva verfügt über die sämtlichen Inhalte der Kantonswebsite als Wissensbasis und ermöglicht es Nutzern, Informationen zu gewünschten Themen abzurufen. Das Abrufen von Informationen funktioniert themen- und bereichsübergreifend, was inhaltlich anspruchsvolle Themen und Prozesse in einfache Schritte herunterbricht und die benötigten Links und Dokumente als Referenzinformationen zusätzlich zur gelieferten Antwort auf die gestellte Anfrage bereitstellt. Alva zählt zum heutigen Zeitpunkt täglich rund 550 Nutzer mit durchschnittlich 1.4 Interaktionen pro Nutzer.

Nach initialen Unterhaltungen ist der Kanton Basel-Stadt einverstanden, das vorgesehene Experiment in der Live-Umgebung von Alva durchzuführen. Die anfallenden Arbeiten zur Integration werden zu je 50% vom Auftraggeber Liip (Liip AG, 2025) und dem Kanton Basel-Stadt getragen. Die benötigte Stimulus-Konzeption und das Survey-Design obliegen in der Verantwortung des Studierenden.

**Tabelle 5**

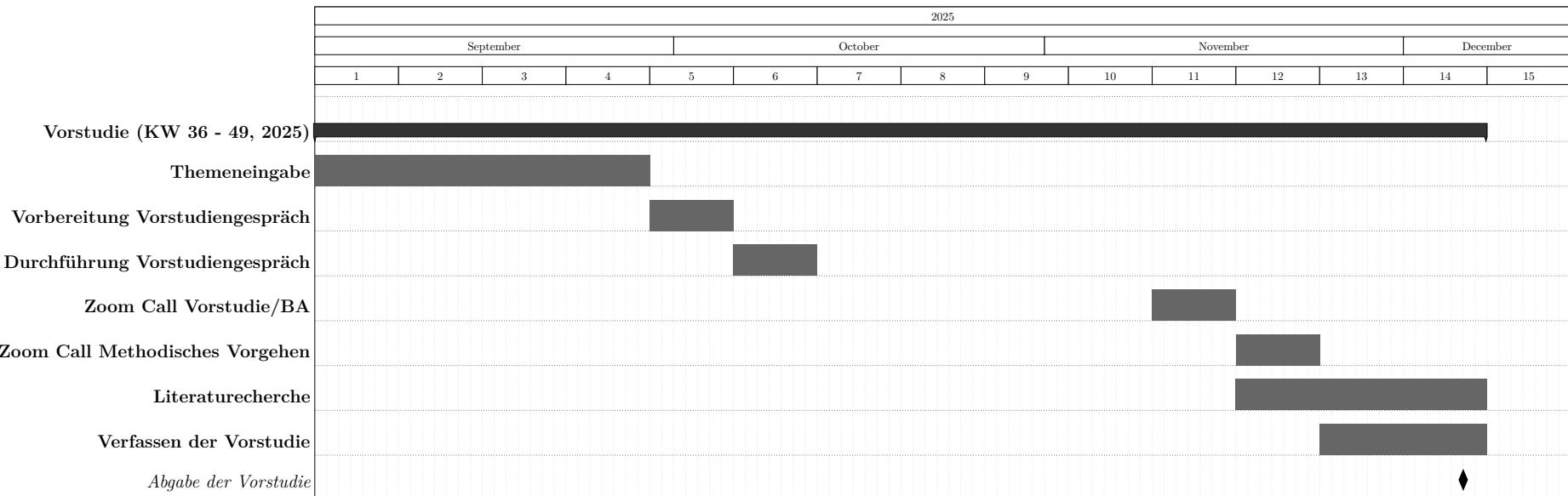
*Meilensteine der Vorstudie und Bachelorarbeit*

Meilenstein	Zeitraum	Beteiligte
Gespräch Machbarkeit intern	Juli 2025	Liip
Gespräch Machbarkeit extern	Oktober 2025	Kanton Basel-Stadt
Entwicklung Anforderungen (Logik & Userflow)	Oktober 2025	Studierender
Schätzung benötigter Arbeiten	November 2025	Product Owner, Frontend Developer
Kommunikation Investment extern	November 2025	Kanton Basel-Stadt
Übereinkunft Investment-Teilung	November 2025	Liip, Kanton Basel-Stadt

### Zielsetzung

## Arbeitsplan

Der folgende Arbeitsplan zeigt die zeitliche Planung der Vorstudie.



**Abbildung 6**

*Gantt-Chart Arbeitsplan*

**Commented Review****Reflection Work**

## Quellenverzeichnis

- Baroni, I., Calegari, G. R., Scandolari, D., & Celino, I. (2022). AI-TAM: a model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications. *Human Computation*, 9(1), 1–21.  
<https://doi.org/10.15346/hc.v9i1.134>
- Cunningham, T., Deming, J. D., Hitzig, Z., Ong, C., Yan Shan, C., & Wadman, K. (2025, September). How People Use ChatGPT. <https://doi.org/10.3386/w34255>
- Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology [Publisher: Management Information Systems Research Center, University of Minnesota]. *MIS Quarterly*, 13(3), 319–340.  
<https://doi.org/10.2307/249008>
- Dolgopolova, I., Li, B., Pirhonen, H., & Roosen, J. (2022). The effect of attribute framing on consumers' attitudes and intentions toward food: A Meta-analysis. *Bio-based and Applied Economics*, 10, 253–264.  
<https://doi.org/10.36253/bae-11511>
- Druckman, J. N. (2001). Evaluating framing effects. *Journal of Economic Psychology*, 22(1), 91–101. [https://doi.org/10.1016/S0167-4870\(00\)00032-5](https://doi.org/10.1016/S0167-4870(00)00032-5)
- Fishbein, M., & Ajzen, I. (1975, Mai). *Belief, attitude, intention and behaviour: An introduction to theory and research* (Bd. 27).
- Freling, T. H., Vincent, L. H., & Henard, D. H. (2014). When not to accentuate the positive: Re-examining valence effects in attribute framing. *Organizational Behavior and Human Decision Processes*, 124(2), 95–109.  
<https://doi.org/10.1016/j.obhdp.2013.12.007>
- Kanton Basel-Stadt. (2025a). *Alva - Der digitale Assistent*. Verfügbar 3. Dezember 2025 unter <https://www.bs.ch>
- Kanton Basel-Stadt. (2025b). *Offizielle Website des Kantons Basel-Stadt*. Verfügbar 3. Dezember 2025 unter <https://www.bs.ch>
- Kim, T., & Song, H. (2022). Communicating the Limitations of AI: The Effect of Message Framing and Ownership on Trust in Artificial Intelligence. *International*

- Journal of Human-Computer Interaction*, 39, 1–11.  
<https://doi.org/10.1080/10447318.2022.2049134>
- Li, J., & Huang, J.-S. (2020). Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technology in Society*, 63, 101410.  
<https://doi.org/10.1016/j.techsoc.2020.101410>
- Liip AG. (2025). *Liip - Digital Progress*. Verfügbar 3. Dezember 2025 unter  
<https://www.liip.ch>
- Tversky, A., & Kahneman, D. (1986, Januar). The Framing of Decisions and the Evaluation of Prospects. In R. Barcan Marcus, G. J. W. Dorn & P. Weingartner (Hrsg.), *Studies in Logic and the Foundations of Mathematics* (S. 503–520, Bd. 114). Elsevier. [https://doi.org/10.1016/S0049-237X\(09\)70710-4](https://doi.org/10.1016/S0049-237X(09)70710-4)

### **Abbildungsverzeichnis**

1	Erweitertes TAM-Modell: Artificial Intelligence-Technology Acceptance Model (Baroni et al., 2022)	5
2	Strukturmodell für AI-TAM, Teil des Strukturgleichungsmodell, inkl. Hypothesen, inkl. Stimulus	11
3	Stimulus-Design: Hohe Sicherheit vs. Tiefe Unsicherheit	15
4	Stimulus-Design: Mittlere Sicherheit vs. Mittlere Unsicherheit	15
5	Übersicht der Stimulus-Designs für das 3x2 Experiment	16
6	Gantt-Chart Arbeitsplan	23

### **Tabellenverzeichnis**

1	Identifizierte latente Konstrukte zum Einsatz von AI-TAM	8
2	Aufgelistete Hypothesen im Rahmen der Bachelor-Arbeit-Vorstudie	9
3	Experiment-Design, vorhandene Experimentalbedingungen	13
4	Experimentelles 3x2 Design: Manipulation von Framing und Accuracy	14
5	Meilensteine der Vorstudie und Bachelorarbeit	22

## Analyse

### **Pretest-Analyse**

#### ***Reliabilitätsanalyse (Cronbach's Alpha)***

Die Reliabilität aller verwendeten Skalen wird mittels Cronbach's Alpha überprüft. Für jedes Konstrukt wird die interne Konsistenz berechnet: AI Output Trust, Perceived Usefulness, Perceived Ease of Use, Behavioral Intention und Collaborative Intention. Bei unzureichender Reliabilität werden Item-Total-Korrelationen analysiert, um problematische Items zu identifizieren. Items mit ungenügender Trennschärfe werden überarbeitet oder eliminiert. Zusätzlich wird die Veränderung des Alpha-Werts bei Itemausschluss berechnet (Alpha if item deleted), um die optimale Itemkombination zu bestimmen.

#### ***Manipulationscheck (t-test unabh. Stichproben)***

Die Wirksamkeit der experimentellen Manipulation wird durch ein t-test für unabhängige Stichproben überprüft. Die Antworten auf die Manipulationscheck-Fragen werden nach Experimentalgruppen aufgeschlüsselt analysiert. Bei der Positiv-Frame-Gruppe wird erwartet, dass Teilnehmende «Konfidenz/Sicherheit» als Darstellungsform angeben, bei der Negativ-Frame-Gruppe «Unsicherheit/Fehlerwahrscheinlichkeit». Zusätzlich wird mittels einfaktorieller ANOVA getestet, ob sich die wahrgenommene Vertrauenswürdigkeit zwischen den Gruppen bereits im Pretest signifikant unterscheidet. Falls die Manipulation zu schwach ist (< 80% Erkennungsrate), werden Anpassungen am Stimulus vorgenommen.

### **Hauptanalyse: Deskriptive Statistik**

#### ***Stichprobenbeschreibung und Gruppenvergleichbarkeit***

Die deskriptive Analyse beginnt mit der Charakterisierung der Stichprobe. Für alle demografischen Variablen (Alter, Geschlecht, Bildung) werden Häufigkeiten, Mittelwerte und Standardabweichungen berichtet, sowohl für die Gesamtstichprobe als auch getrennt nach Experimentalgruppen. Die Vergleichbarkeit der randomisierten Gruppen wird mittels unabhängigem t-test und einfaktorieller ANOVAs (kontinuierliche

Variablen) überprüft. Signifikante Unterschiede würden auf Randomisierungsprobleme hinweisen und müssten in späteren Analysen als Kovariaten berücksichtigt werden. Zusätzlich werden die Verteilungen der AI-Vorerfahrung und Nutzungserfahrung zwischen den Gruppen verglichen (potenzielle Konfundierungsvariablen).

### ***Skaleneigenschaften und Verteilungsanalysen***

Für alle erhobenen Konstrukte werden Mittelwerte, Standardabweichungen, Schiefe berichtet und mittels Histogrammen visualisiert. Q-Q-Plots visualisieren Abweichungen von der Normalverteilung. Die finale Reliabilität der Skalen wird erneut mit Cronbach's Alpha bestimmt und mit den Pretest-Werten verglichen. Interkorrelationen zwischen allen Konstrukten werden in einer Korrelationsmatrix dargestellt, um erste Hinweise auf Zusammenhänge zu erhalten und Multikollinearität zu identifizieren. Die deskriptiven Statistiken der Trust-Ratings während der Interaktion werden separat für jede Alva-Antwort berichtet, um mögliche Veränderungen über die Interaktionen hinweg zu identifizieren.

### ***Manipulationscheck***

Mittels einfaktorieller ANOVA wird getestet, ob sich die wahrgenommene Vertrauenswürdigkeit zwischen den Gruppen signifikant unterscheidet.

### ***Nutzungsverhalten und Systemvariablen***

Die während der LLM-Interaktion automatisch erfassten Variablen werden deskriptiv ausgewertet. Die durchschnittliche Anzahl der Interaktionen pro Person und Gruppe wird berichtet, ebenso die Verteilung der tatsächlichen Accuracy Scores. Die Session-Dauer wird analysiert, um Hinweise auf Engagement oder Frustration zu erhalten. Die Komplexität der gestellten Fragen wird kategorisiert (einfach/mittel/komplex) und deren Verteilung zwischen den Gruppen verglichen. Diese Variablen dienen später als Kontrollvariablen in den Hauptanalysen und ermöglichen es, die ökologische Validität der Ergebnisse zu bewerten.

## **Hauptanalyse: Strukturgleichungsmodellierung (SEM)**

Das verwendete AI-TAM-Modell wird mittels Strukturgleichungsmodell (SEM) analysiert (Bollen, 1989). Die Methode ermöglicht die simultane Schätzung aller Hypothesen (H1–H9) und berücksichtigt dabei die angenommenen Abhängigkeiten zwischen den Konstrukten.

Das Modell umfasst zwei Komponenten: Das Messmodell spezifiziert die Beziehungen zwischen latenten Konstrukten und ihren manifesten Indikatoren. Das Strukturmodell (Abbildung 2) bildet die theoretischen Pfade zwischen den Konstrukten ab. Als exogene Variablen fungieren die Framing-Manipulation (kodiert durch zwei Dummy-Variablen: Dummy\_Pos = 1 für positives Framing, 0 sonst; Dummy\_Neg = 1 für negatives Framing, 0 sonst; mit der Kontrollgruppe ohne Score-Anzeige als Referenzkategorie), der Accuracy Score (ACTS) sowie die Vertrautheit mit Technologie (FAM-TEC). Als endogene latente Variablen werden AI Output Trust (XAIT), Perceived Usefulness (PUF), Perceived Ease of Use (EOU), Behavioral Intention (BI) und Collaborative Intention (CI) spezifiziert. Tabelle 2 zeigt die Zuordnung der Hypothesen zu den entsprechenden Modellpfaden.

### ***Parameterschätzung und Modell-Fit***

Die Parameterschätzung erfolgt mittels Maximum-Likelihood-Verfahren (Bollen, 1989; Kano et al., 1997). Zur Beurteilung des Modelfits werden mehrere Fit-Indizes herangezogen:

- Comparative Fit Index (CFI): Cutoff  $\geq .95$  für sehr guten,  $\geq .90$  für akzeptablen Fit (Bentler, 1990)
- Tucker-Lewis Index (TLI): Cutoff  $\geq .95$  für sehr guten,  $\geq .90$  für akzeptablen Fit (Tucker & Lewis, 1973)
- Root Mean Square Error of Approximation (RMSEA): Cutoff  $\leq .06$  für sehr guten,  $\leq .08$  für akzeptablen Fit (Browne & Cudeck, 1992; Hu & Bentler, 1999; Steiger & Lind, 1980)

- Standardized Root Mean Square Residual (SRMR): Cutoff  $\leq .08$  (Bentler, 1995; Jöreskog & Sörbom, 1981)

### ***Evaluation des Messmodells***

Zunächst wird das Messmodell mittels konfirmatorischer Faktorenanalyse (CFA) geprüft. Dabei werden die fünf latenten Konstrukte (XAIT, PUF, EOU, BI, CI) ohne strukturelle Pfade analysiert, jedoch mit freien Kovarianzen zwischen allen Konstrukten. Dieses zweistufige Vorgehen (Anderson & Gerbing, 1988) ermöglicht die separate Evaluation der Messqualität, bevor die theoretischen Hypothesen im Strukturmodell getestet werden. Dazu werden drei Kriterien betrachtet.

- Faktorladungen: Items mit standardisierten Ladungen  $< .60$  werden für eine theoriegeleitete Modellmodifikation in Betracht gezogen (Gäde et al., 2020)
- Reliabilität: Für jedes Konstrukt wird Cronbach's Alpha berechnet ( $\alpha \geq .80$  als akzeptabel; (Nunnally & Bernstein, 1994))
- Validität: Die Korrelationen zwischen den latenten Konstrukten sollten hoch sein ( $< .85$ )

### ***Strukturmodell und Hypothesentests***

Nach Bestätigung des Messmodells wird das vollständige Strukturmodell geschätzt. Die strukturellen Pfade testen die Hypothesen H2–H8 (AI-TAM- und TAM-Standardbeziehungen) sowie H1a und H1b (Framing-Effekte auf XAIT). Alle Pfade werden simultan geschätzt, wodurch die gegenseitigen Abhängigkeiten zwischen den endogenen Variablen berücksichtigt werden. Für jede endogene Variable wird  $R^2$  berichtet, welches den durch die Prädiktoren erklärten Varianzanteil angibt. Die Signifikanz der Pfadkoeffizienten wird mittels z-Test geprüft ( $\alpha = .05$ , zweiseitig).

### ***Indirekte Effekte und Mediation (H9)***

Die Mediationshypothese (H9) besagt, dass der Effekt des Framings auf «Behavioral Intention» durch «Explainable AI Trust» und die nachgelagerten

TAM-Konstrukte vermittelt wird. Im vorliegenden Modell existieren multiple indirekte Pfade von Framing zu BI:

- Via PUF: Framing → XAIT → PUF → BI
- Via EOU: Framing → XAIT → EOU → BI
- Via EOU und PUF (seriell): Framing → XAIT → EOU → PUF → BI

Diese indirekten Effekte werden simultan im SEM berechnet. Die Signifikanz wird mittels Bias-Corrected Bootstrap-Verfahren mit 5000 Ziehungen und 95%-Konfidenzintervallen getestet. Separate Effekte werden für beide Framing-Bedingungen (Dummy\_Pos, Dummy\_Neg) relativ zur Kontrollgruppe berechnet.

Berichtet werden:

- Die spezifischen indirekten Effekte über jeden einzelnen Mediationspfad
- Der totale indirekte Effekt (Summe aller indirekten Pfade)
- Der totale Effekt (Summe aus direktem und indirektem Effekt, sofern ein direkter Pfad Framing → BI spezifiziert wird)

Eine vollständige Mediation liegt vor, wenn kein signifikanter direkter Effekt von Framing auf BI besteht; eine partielle Mediation liegt vor, wenn zusätzlich zu den indirekten Effekten ein signifikanter direkter Effekt verbleibt (Baron & Kenny, 1986).

### ***Kontrolle für Störvariablen***

Demografische Variablen (Alter, Geschlecht, Bildungsabschlusswerden als Kovariaten im Modell berücksichtigt.