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 mediierte 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

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 («Behavourial 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, 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, definiert als «the degree to which an individual believes that using a particular system would enhance his or her job performance», sowie Perceived Ease of Use, verstanden als «the degree to which an individual believes that using a particular system would be free of physical and mental effort».

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. Das im AI-TAM verwendete Vertrauenskonstrukt entstammt der Forschung von Hoffman et al. (2019) und erfasst, inwieweit Nutzer den Ergebnissen eines KI-Systems vertrauen. Ergänzend misst die Kollaborationsabsicht die Bereitschaft, aktiv an der Weiterentwicklung der KI mitzuwirken.

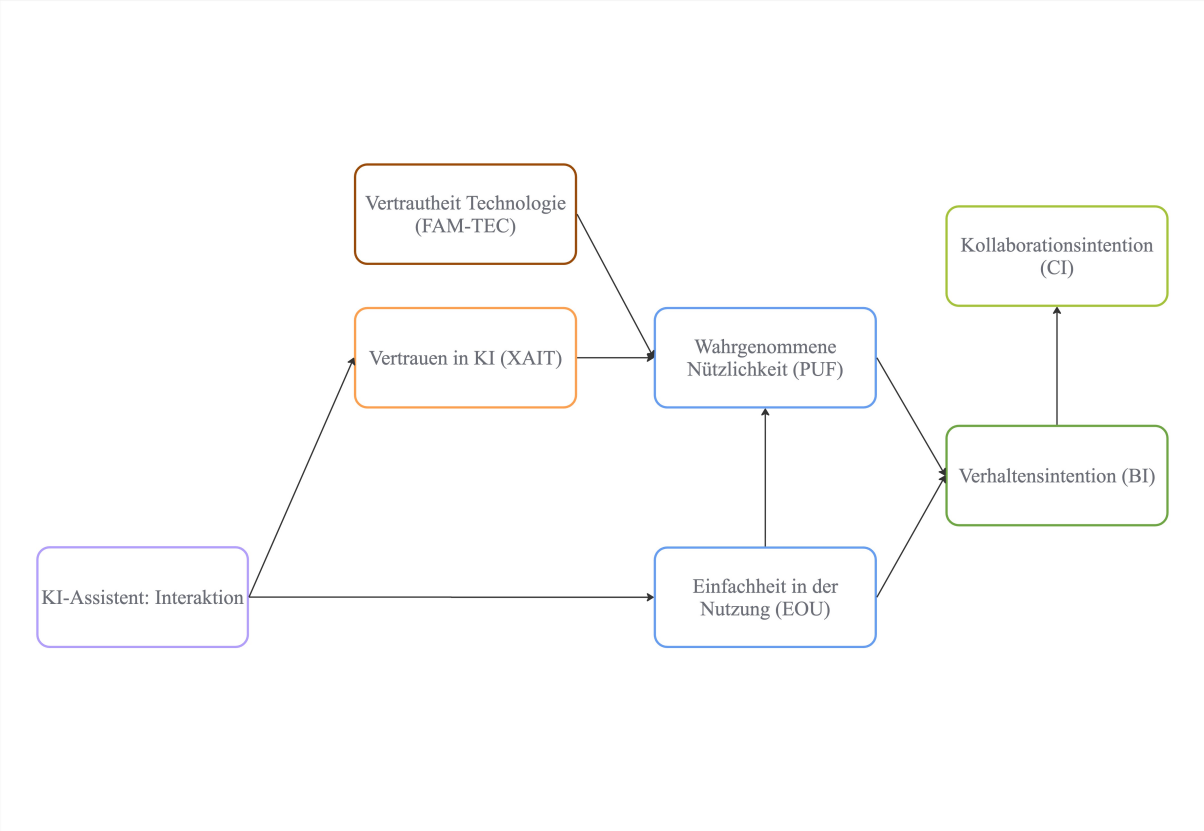

Abbildung 1

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

Framing-Theorie als Stimulus-Konzept. Die Verbindung des AI-TAM mit dem Konzept des Framings eröffnet neue Forschungsperspektiven. Besonders das Attribute Framing könnte das XAIT-Konstrukt beeinflussen: Die Präsentation von

KI-Fähigkeiten mit unterschiedlichen Konfidenz-Stufen (hoch, mittel, tief) dürfte das Vertrauen in die KI direkt verändern. Gemäss dem AI-TAM-Modell beeinflusst dieser Faktor wiederum die Nutzungsabsicht. Für Experimente mit Large Language Models bedeutet dies, dass die Art der Leistungsdarstellung die Nutzerakzeptanz beeinflussen könnte. Das AI-TAM bietet hierbei den methodischen Rahmen, um diese Effekte präzise zu messen. Im Folgenden wird das Konzept des Attribute Framings näher erläutert.

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.

Freling et al. (2014) führten eine umfassende Meta-Analyse von 107 Studien zum Thema Attribute Framing durch. Attribute Framing bezeichnet die Darstellung identischer Informationen in unterschiedlicher Formulierung – beispielsweise „80% Erfolgsrate"versus „20% Misserfolgsrate". Die etablierte Forschungsmeinung ging davon aus, dass positive Formulierungen grundsätzlich wirksamer sind als negative.

Die Autoren erweiterten diese Perspektive, indem sie die Construal Level Theory (CLT) als theoretischen Rahmen anwendeten. Diese Theorie unterscheidet zwischen abstraktem Denken (z.B. übergeordnete Ziele wie „erfolgreich sein") und konkretem Denken (z.B. spezifische Handlungen wie „heute den Bericht fertigstellen"). Gleichzeitig berücksichtigt die Theorie die psychologische Distanz – also wie nah oder fern sich eine Person einem Ereignis fühlt, sei es zeitlich, räumlich, sozial oder in Bezug auf die Wahrscheinlichkeit des Eintretens.

Die zentrale Erkenntnis der Meta-Analyse: Die Wirksamkeit eines Frames hängt nicht allein von seiner positiven oder negativen Formulierung ab, sondern von der

Passung zwischen Abstraktionsniveau und psychologischer Distanz. Konkret bedeutet dies: Abstrakte Botschaften wirken besser bei psychologisch entfernten Ereignissen (z.B. Entscheidungen für die ferne Zukunft), während konkrete Botschaften bei psychologisch nahen Ereignissen effektiver sind. Positive Formulierungen fördern dabei eher abstraktes Denken, negative Formulierungen eher konkretes Denken. Die Autoren schlussfolgern daher, dass nicht das Vorzeichen der Botschaft entscheidend ist, sondern die Kongruenz zwischen der Darstellungsweise und der wahrgenommenen Nähe zum Thema (Freling et al., 2014).

Attribute Framing nach I. P. Levin und 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) fanden bei Lebensmittelentscheidungen differenzierte Effekte: Gain-Frames erzeugten positivere Einstellungen, jedoch keinen signifikanten Effekt auf Kaufintentionen. Der Framing-Effekt variiert somit je nach abhängiger Variable. Für die KI-Akzeptanz ist Attribute Framing relevant, da KI-Systeme durch unterschiedliche Konfidenz-Darstellungen charakterisiert werden können (vgl. Kim & Song, 2022; I. Levin & Schneider, 1998). Das AI-TAM bietet den Rahmen, um diese Effekte auf Vertrauen und Nutzungsabsicht zu untersuchen.

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 nach der Nutzung der KI-Assistenz erhoben. Die verwendeten Konstrukte basieren auf dem AI-TAM von Baroni et al. (2022) und wurden aus drei Quellen adaptiert: Ibrahim et al. (2025) für die Kernkonstrukte wahrgenommene Nützlichkeit, Einfachheit der Nutzung, Verhaltensintention und Vertrauen in KI; Topsakal (2025) für die technologische Vorerfahrung; sowie Grassi et al. (2022) für die Kollaborationsintention. Alle Items werden auf einer 5-stufigen Likert-Skala erhoben.

Hypothesenübersicht

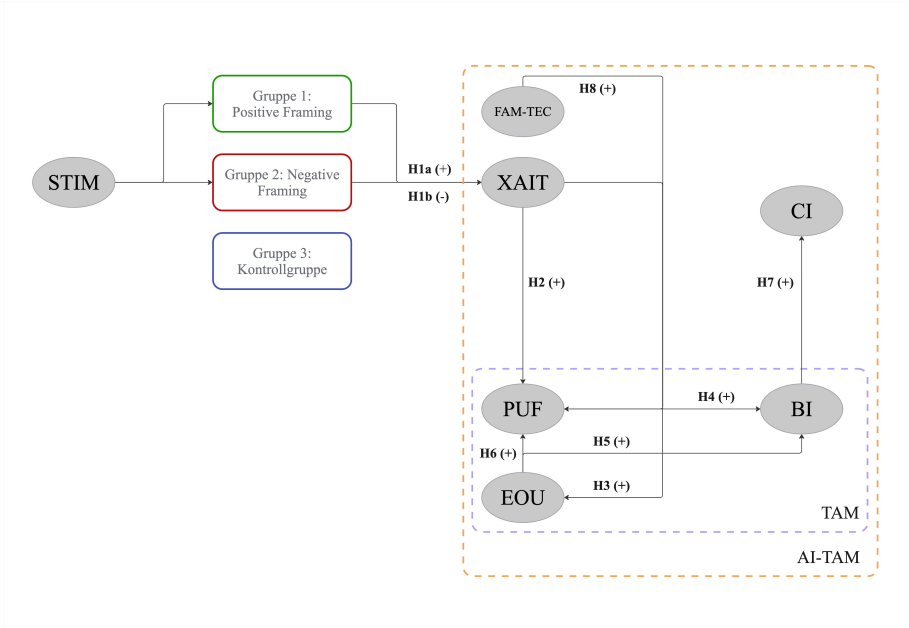

Abbildung 2

Hypothesenmodell

Tabelle 3

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

Forschungsdesign

Das Experiment untersucht den Einfluss von Framing bezüglich Sicherheit und Unsicherheit auf das Vertrauen in KI-Systeme. In einem 3x2 Between-Subjects-Design wird die Darstellung von Konfidenzwerten (Sicherheit vs. Unsicherheit) bei variierenden Accuracy-Scores (hoch, mittel, niedrig) manipuliert. Daraus ergeben sich sechs Experimentalgruppen sowie eine Kontrollgruppe ohne Score-Anzeige. Die experimentelle Manipulation erfolgt während der realen Interaktion mit einem KI-Assistenten.

Experimentelles Design

Das Untersuchungsdesign entspricht einem 3x2 faktoriellen Between-Subjects-Design. Die erste unabhängige Variable (Framing) variiert die

Darstellung als Sicherheit versus Unsicherheit, die zweite unabhängige Variable (Accuracy) variiert den Konfidenzwert in drei Stufen (hoch, mittel, niedrig). Dieses Design ermöglicht die Untersuchung von Haupt- und Interaktionseffekten beider Faktoren auf die abhängigen Variablen

Stimulus-Konzept

Der Stimulus besteht aus der visuellen und textlichen Darstellung einer Sicherheits- bzw. Unsicherheitsanzeige, die direkt nach jeder LLM-Antwort entsprechend der zugewiesenen Stimulusgruppe eingeblendet wird. Die Manipulation erfolgt in Echtzeit während der natürlichen Interaktion mit dem digitalen Assistenten.

Das Stimulus-Design setzt sich aus zwei Dimensionen zusammen. Die erste Dimension betrifft die Valenz der Darstellung: Die Anzeige wird entweder positiv als „Sicherheitöder negativ als „Unsicherheit"gerahmt. Die zweite Dimension umfasst die Ausprägungsstufe, wobei die angezeigte Sicherheit bzw. Unsicherheit in drei Stufen variiert – hoch, mittel und niedrig. Aus der Kombination dieser beiden Dimensionen ergibt sich ein 2×3-Design mit insgesamt sechs unterschiedlichen Stimulus-Bedingungen, die jeweils eine spezifische visuelle und textliche Gestaltung aufweisen.

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.

Ablauf Experiment

Das geplante Experiment findet in drei Phasen statt. In der ersten Phase werden die Nutzenden über das Experiment informiert und können sich für oder gegen eine Teilnahme entscheiden. In Phase 2 steht die Interaktion mit dem Chatbot Alva im Zentrum. In Phase 3 werden die Nutzenden aufgefordert, die dazugehörige Umfrage auszufüllen und das Experiment abzuschliessen. Es werden keine Daten im Vorfeld (Phase 1) oder während der Interaktion (Phase 2) erhoben, um die Abbruchrate zu minimieren und eine hohe Abschlussrate zu fördern.

Methodische Einordnung des Forschungsdesigns

Das vorliegende Forschungsdesign verbindet ein kontrolliertes Experiment mit einer Felderhebung im realen Nutzungskontext. Die Wahl dieser Methode orientiert sich an der Fragestellung und dem untersuchten Gegenstandsbereich (vgl. Kelle, 2022, S. 174f.).

Die experimentelle Manipulation im Between-Subject-Design mit randomisierter Zuweisung zu den Experimentalbedingungen gewährleistet die interne Validität. Durch die Randomisierung wird sichergestellt, dass beobachtete Unterschiede in den abhängigen Variablen auf die experimentelle Manipulation (Framing und Konfidenz-Level) zurückgeführt werden können. Die Einbettung des Experiments in die tatsächliche Alva-Nutzung erhöht die ökologische Validität gegenüber rein laborbasierten oder szenariobasierten Designs. Proband*innen stellen eigene Fragen im realistischen Anwendungskontext, anstatt auf vorgegebene Szenarien zu reagieren. Die standardisierte Post-Befragung mittels validierter Skalen aus dem AI-TAM-Modell (Baroni et al., 2022) ermöglicht die reliable Messung der latenten Konstrukte.

Methodenintegration

Die Integration der Datenquellen erfolgt auf Analyseebene: Die experimentelle Gruppenzugehörigkeit (Framing und Konfidenz-Level) wird mit den Befragungsdaten (Vertrauen, TAM-Konstrukte) in einem gemeinsamen Datensatz zusammengeführt. Diese Integration ermöglicht die Analyse von Haupteffekten des Framings (Sicherheit vs. Unsicherheit) und des Konfidenz-Levels (hoch, mittel, tief) auf das Vertrauen, von Interaktionseffekten zwischen beiden Faktoren sowie von Pfadbeziehungen zwischen Vertrauen und den TAM-Konstrukten gemäss dem AI-TAM-Modell.

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.

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» (Alva Team, 2025) dient als zentraler digitaler Assistent bei der Bedienung und Navigation der Website des Kantons Basel-Stadt (Kanton Basel-Stadt, 2025). 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, 2025) und dem Kanton Basel-Stadt getragen. Die benötigte Stimulus-Konzeption und

das Survey-Design obliegen in der Verantwortung des Studierenden.

Zielsetzung

Tabelle 1

Identifizierte latente Konstrukte zum Einsatz von AI-TAM

Abkürzung	**Name**	**Definition**	**Quelle**
XAIT	Explainable AI Trust, Vertrauen in KI	Vertrauen in die generierte Antwort und die LLM-Lösung	(Ibrahim et al., 2025)
BI	Behavioral Intention, Verhaltensintention	Verhaltensintention die LLM-Lösung zu Nutzen	(Ibrahim et al., 2025)
CI	Collaborative Intention, Kollaborationsintention	Kollaborationsintention zur digitalen Assistenz	(Grassi et al., 2022)
PUF	Perceived Usefulness, Wahrgenommene Nützlichkeit	Wahrgenommene Nützlichkeit der generierten Antwort	(Ibrahim et al., 2025)
EOU	Ease of Use, Einfachheit der Nutzung	Einfachheit in der Nutzung der KI-Applikation	(Ibrahim et al., 2025)
FAM-TEC	Familiarity with Technology, Technologische Vorerfahrung	Vertrautheit in der Nutzung von KI-Technologie	(Topsakal, 2025)

Tabelle 5

Experiment-Design, vorhandene Experimentalbedingungen

Bedingung	**Gruppe**	**Manipulation**	**Beispiel**
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 dargestellt	Antwortunsicherheit: 20% oder «Diese Antwort hat eine Fehlerwahrscheinlichkeit von 20%»
Kontrollgruppe	Gruppe 3	Kein Score wird angezeigt (Status Quo)	-

Tabelle 7

Experimentelles 3x2 Design: Manipulation von Framing und Accuracy

	Framing (Unabhängige Variable 1)	
Accuracy (UV 2)	**Positiver Frame**	**Negativer Frame**
Hoch	Sicherheit: Hoch	Unsicherheit: Tief
Mittel	Sicherheit: Mittel	Unsicherheit: Mittel
Niedrig	Sicherheit: Tief	Unsicherheit: Hoch

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

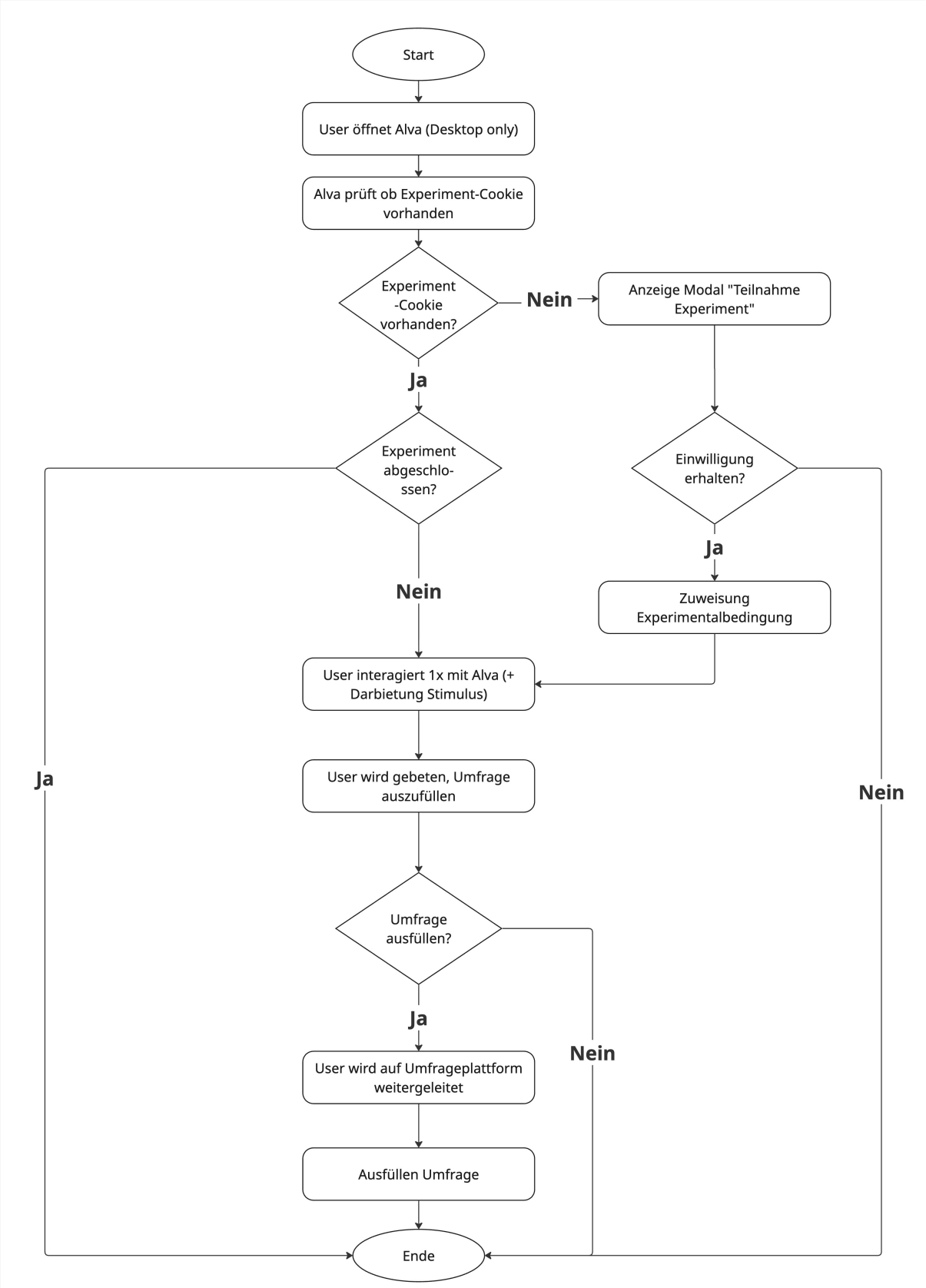

Abbildung 3

Ablauf des Experiments in drei Phasen

Tabelle 9

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

Arbeitsplan

Der folgende Arbeitsplan zeigt die zeitliche Planung der Vorstudie.

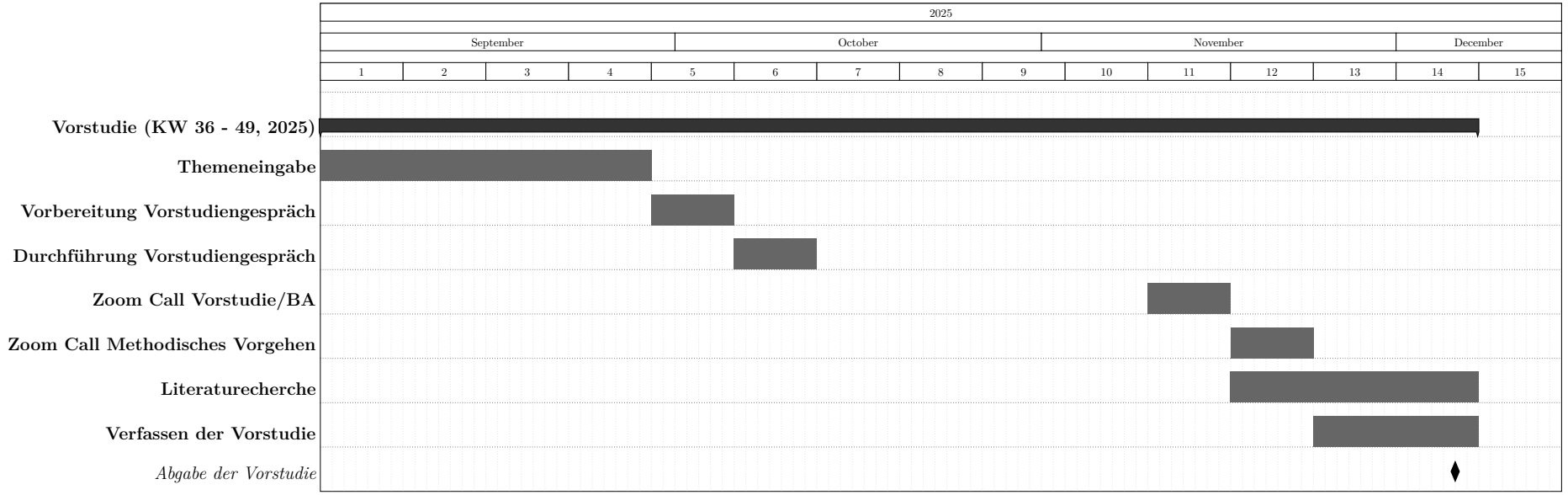

Abbildung 4

Gantt-Chart Arbeitsplan

Commented Review

Reflection Work

Quellenverzeichnis

Alva Team. (2025). Alva: AI Assistant [Internal Project Documentation].

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

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

Grassi, L., Recchiuto, C., & Sgorbissa, A. (2022). Knowledge-Grounded Dialogue Flow Management for Social Robots and Conversational Agents. *International Journal of Social Robotics*, *14*. https://doi.org/10.1007/s12369-022-00868-z

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, D. (2019). Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608*.

Ibrahim, F., Münscher, J.-C., Daseking, M., & Telle, N.-T. (2025). The technology acceptance model and adopter type analysis in the context of artificial intelligence [Publisher: Frontiers]. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1496518

Kanton Basel-Stadt. (2025). Kanton Basel-Stadt. https://www.bs.ch

Kelle, U. (2022). Mixed Methods. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 173–185). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-37985-8_12

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

Levin, I., & Schneider, S. (1998). All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects, *Organizational Behavior and Human Decision Processes*, *76*, 149–188. https://doi.org/10.1006/obhd.1998.2804

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research*, *15*(3), 374–378.

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. (2025). Liip AG. https://www.liip.ch

Topsakal, Y. (2025). How Familiarity, Ease of Use, Usefulness, and Trust Influence the Acceptance of Generative Artificial Intelligence (AI)-Assisted Travel Planning [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2024.2426044]. *International Journal of Human–Computer Interaction*, *41*(15), 9478–9491. https://doi.org/10.1080/10447318.2024.2426044

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

Abbildungsverzeichnis

Tabellenverzeichnis