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

Inhaltsverzeichnis

1 Einleitung

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 (OpenAI, 2025), Claude von Anthropic (Anthropic, 2025) und Gemini von Google (Inc., 2025) 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 1989 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 und Gemini weisen vor- sowie während der Nutzung ausdrücklich darauf hin, dass ihre Modelle 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.

1.1 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.

1.2 Zielsetzung

Die vorliegende Bachelorarbeit verfolgt zwei zentrale Ziele: Erstens soll empirisch untersucht werden, wie unterschiedliche Darstellungsformen von KI-Konfidenzwerten (positives vs. negatives Framing) das Nutzervertrauen in LLM-basierte Assistenzsysteme beeinflussen (Kim & Song, 2022; Levin & Schneider, 1998). Zweitens soll die Anwendbarkeit des Artificial Intelligence Technology Acceptance Model (AI-TAM) im Kontext eines öffentlichen KI-Assistenten validiert werden, insbesondere hinsichtlich der Mediationsrolle von Vertrauen zwischen Transparenzkommunikation und Nutzungsintention (Baroni et al., 2022).

1.3 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.

Tabelle 1

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

1.4 Arbeitsplan

Der folgende Arbeitsplan zeigt die zeitliche Planung der Vorstudie.

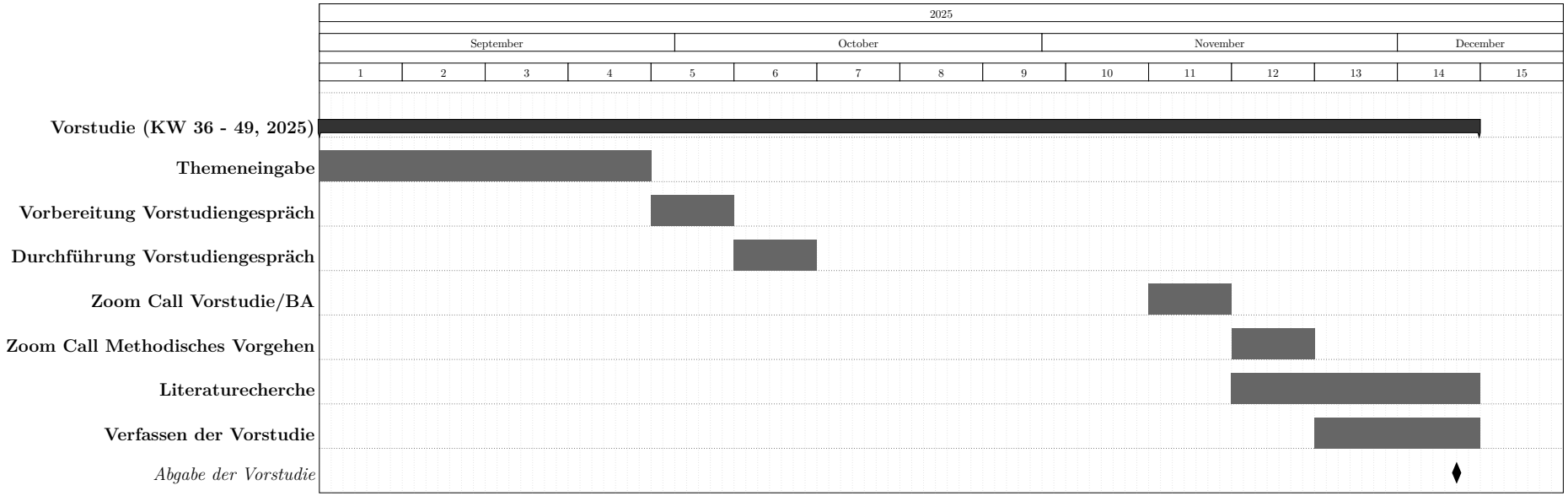

Abbildung 1

Gantt-Chart Arbeitsplan

2 Theoretische Einbettung

Modelltheoretisch knüpft die vorliegende Arbeit an frühere Studien in den Bereichen Vertrauen in künstliche Intelligenz 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).

2.1 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. 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».

2.2 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.

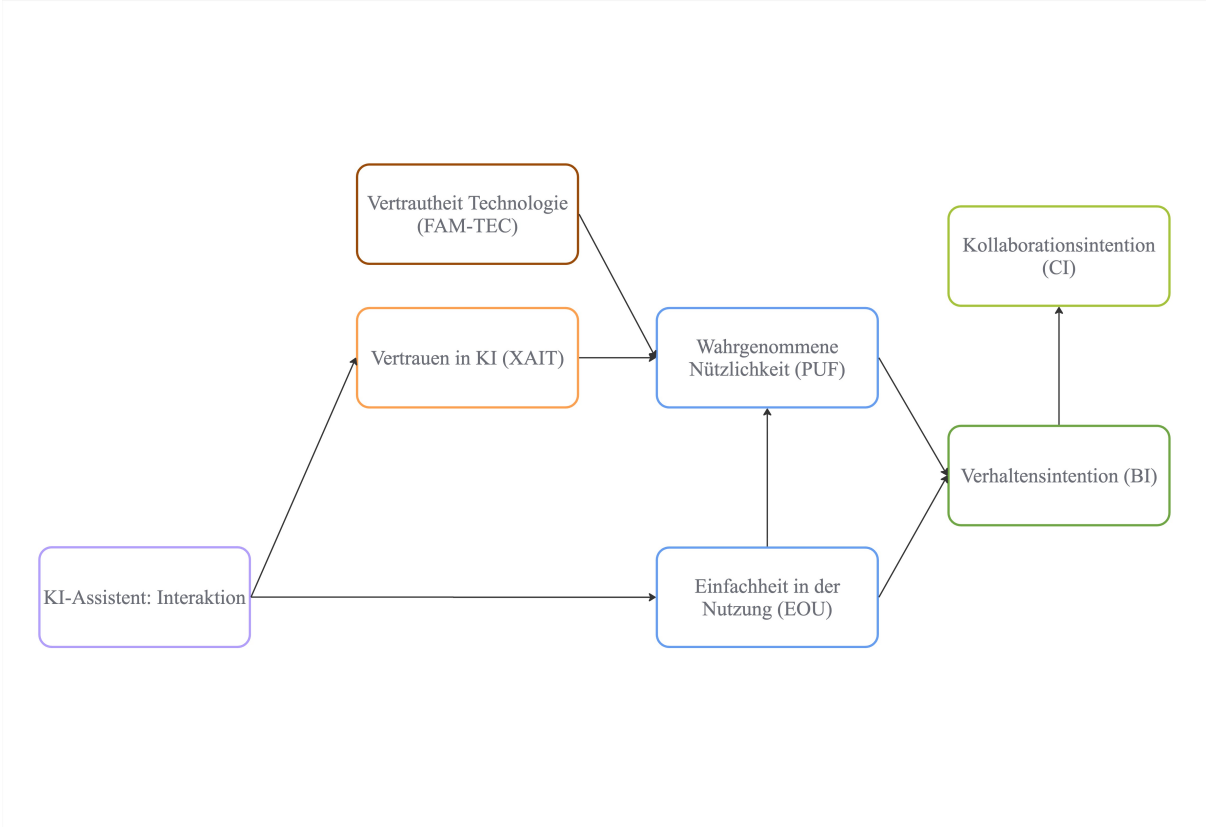

Abbildung 2

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

2.3 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 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, dass nicht das Vorzeichen der Botschaft entscheidend ist, sondern die Kongruenz zwischen der Darstellungsweise und der wahrgenommenen Nähe zum Thema (Freling et al., 2014).

2.4 Attribute Framing

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. Dieser Befund Befund, der durch ein Averaging-Modell erklärt wird, bei dem zusätzliche Informationsquellen den Einfluss einzelner Frames reduzieren. Dolgopolova et al. (2022) fanden bei Lebensmittelentscheidungen weitere Effekte: Positive–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; Levin & Schneider, 1998). Das AI-TAM bietet den Rahmen, um diese Effekte auf Vertrauen und Nutzungsabsicht zu untersuchen.

2.5 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.

2.6 Hypothesenübersicht

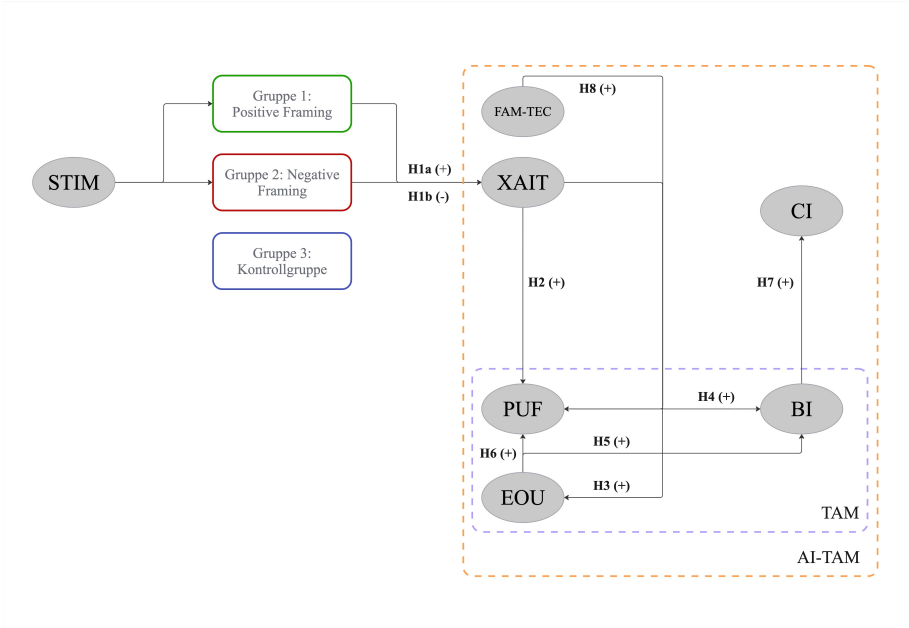

Abbildung 3

Hypothesenmodell

Tabelle 3

Aufgelistete Hypothesen im Rahmen der Bachelor-Arbeit-Vorstudie

Hypothese	Pfad	Richtung	Theorie
H1a	Stimulus Positiv → XAIT	+	Attribute Frame
H1b	Stimulus Negativ → 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

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

Tabelle 5

Ausformulierte Hypothesen

Bereich	Nr.	Hypothese
Framing-Hypothesen		
Framing	H1a	Die Darstellung als Sicherheit (positiver Frame) führt zu einem höheren Vertrauen in künstliche Intelligenz.
	H1b	Die Darstellung als Unsicherheit (negativer Frame) führt zu einem niedrigeren Vertrauen in künstliche Intelligenz.
AI-TAM-Hypothesen		
AI-TAM	H2	Vertrauen in künstliche Intelligenz hat einen positiven Einfluss auf die wahrgenommene Nützlichkeit.
	H3	Vertrauen in künstliche Intelligenz hat einen positiven Einfluss auf die wahrgenommene Einfachheit in der Nutzung.
TAM-Hypothesen		
TAM	H4	Die wahrgenommene Nützlichkeit hat einen positiven Einfluss auf die Nutzungsintention.
	H5	Die wahrgenommene Einfachheit in der Nutzung hat einen positiven Einfluss auf die Nutzungsintention.
	H6	Die wahrgenommene Einfachheit in der Nutzung hat einen positiven Einfluss auf die wahrgenommene Nützlichkeit.
	H7	Die Nutzungsintention hat einen positiven Einfluss auf die Kollaborationsintention.
	H8	Die Vertrautheit mit Technologie hat einen positiven Einfluss auf die wahrgenommene Nützlichkeit.

3 Forschungsfrage

Aus der Auseinandersetzung mit der Literatur und den bestehenden Modellen zur Technologieakzeptanz ergibt sich folgende übergeordnete Forschungsfrage:

Wie beeinflusst das Framing von Konfidenzangaben einer LLM-basierten Applikation das Vertrauen der Nutzer und deren Nutzungsabsicht?

4 Forschungsdesign

Das Experiment untersucht den Einfluss von Framing bezüglich Sicherheit und Unsicherheit auf das Vertrauen in KI-gestützte 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 Konfidenzwert-Anzeige. Die experimentelle Manipulation erfolgt während der realen Interaktion mit einem KI-Assistenten.

4.1 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 (Konfidenzwert) variiert den Konfidenzwert in drei Stufen (hoch, mittel, niedrig).

Tabelle 7

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)	-

4.2 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 (Basel-Stadt, 2025).

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.

Tabelle 9

Experimentelles 3x2 Design: Manipulation von Framing und Konfidenz

	Framing (UV 1)	
Konfidenz (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.

4.2.1 Treatment Check Stimulus

Als Treatment Check 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.

4.3 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.

4.3.1 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 Effekten des Framings (Sicherheit vs. Unsicherheit) und des Konfidenz-Levels (hoch, mittel, tief) auf das Vertrauen.

4.4 Abgrenzung des Forschungsdesigns

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

Tabelle 11

Inhaltliche Abgrenzung des Forschungsdesigns

Aspekt	**Fokus dieser Studie**	**Abgrenzung**
Framing-Typ	Attribute Framing	Risky Choice Framing, Goal Framing
Zeithorizont	Nutzungsintention (einmalige Messung)	Tatsächliche Systemnutzung, Langzeiteffekte
Transparenzmechanismen	Konfidenz-Darstellung	Erklärungen, Quellenangaben, Visualisierungen
Anwendungskontext	Verwaltungskontext (Kanton Basel-Stadt)	Medizinische, kreative oder andere Bereiche

Tabelle 13

Methodische Abgrenzung des Forschungsdesigns

Aspekt	**Fokus dieser Studie**	**Abgrenzung**
Studiendesign	Between-Subjects, 3x2 faktoriell	Within-Subjects
Datenerhebung	Quantitativ (standardisierte Skalen)	Qualitative Vertiefungen
Konfidenz-Level	Experimentell manipuliert	Natürliche Variation

4.5 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.

5 Selbstreflexion und Ausblick

5.1 Selbstreflexion

Die bisherige Arbeit an meiner Bachelor-Thesis bedeutete für mich einen grossen Wissensgewinn in einem Themenfeld, das mich seit Längerem fasziniert. Von Anfang an war mir klar, dass ich ein Experiment durchführen wollte. Die Möglichkeit, eigene Hypothesen empirisch zu prüfen, reizte mich besonders. Dabei konnte ich meinen Interessen folgen: Künstliche Intelligenz, Mensch-Maschine-Interaktion und die psychologischen Faktoren, die unsere Wahrnehmung von Technologie beeinflussen.

Das Ausarbeiten des Forschungsdesigns war eine spannende Erfahrung. Zum ersten Mal hatte ich die Gelegenheit, ein Design so auszugestalten, wie ich es mir vorstellte. Von der Definition der Forschungsfrage über die Hypothesenbildung bis zur Operationalisierung der Konstrukte war dieser Prozess lehrreich und zeigte mir, wie

viele Entscheidungen in einem scheinbar einfachen experimentellen Setup stecken.

Besonders zufrieden bin ich mit meinem Betreuungssetup. Mein Dozent erweist sich als wertvoller Sparringpartner, dessen Erfahrung mir half, mein Design zu schärfen und methodische Fallstricke zu vermeiden. Auf Seiten des Praxispartners darf ich mit einer äusserst versierten Ansprechpartnerin zusammenarbeiten, die das Projekt von Beginn an unterstützte und die nötigen Rahmenbedingungen schuf.

Was mich besonders freut: Das Experiment wird in einer realen Umgebung durchgeführt. Anstatt einer simulierten Trockenübung können echte Nutzende des KI-Assistenten Alva an der Studie teilnehmen. Dies erhöht nicht nur die externe Validität der Ergebnisse, sondern gibt der Arbeit auch eine praktische Relevanz, die über den akademischen Kontext hinausgeht.

5.2 Weiteres Vorgehen

Als nächster Schritt steht die Operationalisierung der Konstrukte an, welche bereits begonnen hat. Parallel dazu werden die User Stories erstellt, welche die technischen Anforderungen für die Integration des Experiments definieren. Sobald die Operationalisierung abgeschlossen ist, folgt ein Pre-Test zur Überprüfung der Stimuli und des Fragebogens. Die Erkenntnisse daraus dienen gegebenenfalls zur Anpassung der Stimuli oder zur Überarbeitung der Items.

Im Januar und Februar 2026 wird die benötigte Experiment-Logik auf der Website des Kantons Basel-Stadt implementiert. Das Experiment selbst ist für den Zeitraum von Mitte Februar bis Mitte April 2026 geplant. Im Anschluss erfolgen die Auswertung der erhobenen Daten und die Erstellung der Bachelor-Arbeit.

6 Quellenverzeichnis

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

Anthropic. (2025, April). Anthropic Claude. Verfügbar 5. Dezember 2025 unter https://claude.ai/

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

Basel-Stadt, K. (2025, August). Alva Chatbot | Kanton Basel-Stadt. Verfügbar 3. Dezember 2025 unter https://www.bs.ch/alva

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

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

Inc., G. (2025, April). Google Gemini. Verfügbar 5. Dezember 2025 unter https://gemini.google.com

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

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

OpenAI. (2025, April). OpenAI ChatGPT. Verfügbar 5. Dezember 2025 unter https://openai.com/

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

7 Kommentiertes Literaturverzeichnis

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

Diese Studie entwickelt und validiert das AI Technology Acceptance Model (AI-TAM), eine Erweiterung des klassischen Technology Acceptance Models (TAM) für KI-gestützte Anwendungen mit Human-in-the-Loop-Mechanismen. Die Autoren integrieren Konstrukte aus der Explainable AI (XAI)-Literatur, insbesondere Nutzervertrauen in KI und wahrgenommene Qualität der KI-Ergebnisse, sowie das neue Konstrukt der Collaborative Intention, das die Bereitschaft der Nutzer misst, aktiv zur Verbesserung von KI-Systemen beizutragen. Die empirische Validierung erfolgte mittels einer Crowdsourcing-Kampagne (N=400) auf der Prolific-Plattform anhand einer App zur Schadensschätzung bei Autounfällen. Die Ergebnisse zeigen, dass XAI-bezogene Faktoren einen starken positiven Effekt auf Nutzungsabsicht, wahrgenommene Nützlichkeit und Benutzerfreundlichkeit haben. Zudem besteht ein signifikanter Zusammenhang zwischen Nutzungsabsicht und Kollaborationsbereitschaft, was die erfolgreiche Implementierung von Human-in-the-Loop-Ansätzen in Endnutzeranwendungen unterstützt.

Der Artikel (48 Zitierungen) ist für die Forschung zur Nutzerakzeptanz von KI-Systemen besonders relevant, da er erstmals ein validiertes Messinstrument für die spezifischen Herausforderungen der Mensch-KI-Kollaboration bereitstellt. Eine Limitation besteht darin, dass das Modell bisher nur an einem spezifischen Anwendungsszenario (Versicherungs-App) getestet wurde. Weitere Validierungen in anderen Kontexten stehen noch aus.

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

Diese Grundlagenstudie entwickelt und validiert das Technology Acceptance Model (TAM), das zwei zentrale Konstrukte zur Vorhersage der Nutzerakzeptanz von Informationstechnologie einführt: Perceived Usefulness (wahrgenommene Nützlichkeit), definiert als der Grad, zu dem eine Person glaubt, dass die Nutzung eines Systems ihre Arbeitsleistung verbessert, und Perceived Ease of Use (wahrgenommene Benutzerfreundlichkeit), der Grad, zu dem die Nutzung als aufwandsfrei empfunden wird. Davis entwickelte Skalen, die in zwei empirischen Studien mit insgesamt 152 Nutzern und vier Anwendungsprogrammen getestet wurden. Die resultierenden 6-Item-Skalen erreichten gute bis sehr gute Reliabilitätswerte (α=.98 für Usefulness, α=.94 für Ease of Use) sowie hohe konvergente, diskriminante und faktorielle Validität. Die Ergebnisse zeigen, dass Perceived Usefulness signifikant stärker mit der Systemnutzung korreliert (r=.63 bzw. r=.85) als Perceived Ease of Use (r=.45 bzw. r=.59). Regressionsanalysen legen nahe, dass Ease of Use primär als kausaler Antezedent von Usefulness wirkt und nicht als paralleler, direkter Prädiktor der Nutzung. Der Artikel (265 Zitierungen) gilt als Grundlagenwerk der Technologieakzeptanzforschung und hat das TAM als eines der meistzitierten Modelle in der Wirtschaftsinformatik etabliert. Die validierten Skalen werden bis heute in zahlreichen Studien zur Nutzerakzeptanz eingesetzt. Eine Limitation besteht darin, dass die Nutzungsmessung auf Selbstberichten basiert und keine objektiven Nutzungsdaten erhoben wurden.

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

Diese Metaanalyse entwickelt eine Typologie zur Unterscheidung dreier grundlegend verschiedener Arten von Valenz-Framing-Effekten, die in der Literatur häufig vermischt wurden: Risky Choice Framing beeinflusst die Risikobereitschaft bei Entscheidungen zwischen sicheren und unsicheren Optionen (z.B. Tversky & Kahnemans „Asian Disease Problem"); Attribute Framing beeinflusst die Bewertung einzelner Objekteigenschaften (z.B. „75% mager"vs. „25% fett"bei Hackfleisch); und Goal Framing beeinflusst die Überzeugungskraft von Botschaften durch Betonung von Gewinnen oder Verlusten (z.B. Brustselbstuntersuchung). Die Autoren analysieren systematisch die unterschiedlichen operationalen Definitionen, abhängigen Variablen und zugrundeliegenden psychologischen Mechanismen jedes Framing-Typs. Während Risky Choice Framing durch Prospect Theory erklärt wird, basiert Attribute Framing auf assoziativem Encoding in valenzkongruenten Gedächtnisstrukturen, und Goal Framing wird durch Verlustaversion und Negativitätsbias erklärt. Die Typologie löst scheinbare Widersprüche in der Literatur auf, etwa warum positive Frames bei Attributbewertungen vorteilhaft sind, während negative Frames bei Überzeugungskommunikation wirksamer sein können. Der Artikel (1.899 Zitierungen) gilt als Standardwerk zur konzeptuellen Differenzierung von Framing-Effekten und hat die methodische Präzision in der Entscheidungsforschung massgeblich beeinflusst. Die Typologie wird bis heute als Orientierungsrahmen für die Einordnung und Gestaltung von Framing-Studien verwendet. Eine Limitation besteht darin, dass die Grenzen zwischen den drei Framing-Typen in komplexen Realsituationen nicht immer trennscharf sind und Mischformen auftreten können.

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

Diese Studie hinterfragt die etablierte Annahme, dass positive Attribut-Frames grundsätzlich wirksamer sind als negative. Die Autoren wenden die Construal Level Theory (CLT) auf die Attribut-Framing-Forschung an und führen zunächst eine Meta-Analyse von 107 publizierten Artikeln (N=88.326 Beobachtungen) durch. Die Erkenntnis ist, dass Framing-Effekte am stärksten sind, wenn Kongruenz zwischen dem Konstrual-Level der Botschaft (abstrakt vs. konkret) und der psychologischen Distanz des Empfängers zum geframten Ereignis besteht. Das vorher erwähnte Folgeexperiment (N=100) bestätigt, dass nicht die Valenz allein, sondern die Übereinstimmung zwischen Konstrual-Level und psychologischer Distanz (temporal, hypothetisch, affektiv, informationell, sozial) die Framing-Effekte treibt. Die Ergebnisse zeigen: Positive Frames wirken besser bei psychologisch distanten Ereignissen, während negative Frames bei psychologisch nahen Ereignissen effektiver sein können.

Der Artikel (52 Zitierungen) leistet einen wichtigen Beitrag zur Framing-Forschung, indem er den Fokus von der reinen Botschaftsgestaltung auf die komplexe Beziehung zwischen Botschaft und Empfänger verlagert. Dies hat praktische Implikationen für die Gestaltung persuasiver Kommunikation in Marketing, Führung und Verhandlungen. Eine Limitation besteht darin, dass die Meta-Analyse auf publizierte Studien beschränkt ist, was potenzielle Publication-Bias-Effekte nicht vollständig ausschliesst.

8 Abbildungsverzeichnis

9 Tabellenverzeichnis

10 Glossar

Tabelle 15

Glossar der verwendeten Abkürzungen

Abkürzung	Begriff Englisch	Begriff Deutsch
LLM	Large Language Model	Grosses Sprachmodell
KI	Artificial Intelligence (AI)	Künstliche Intelligenz
TAM	Technology Acceptance Model	Technologieakzeptanzmodell
AI-TAM	AI Technology Acceptance Model	KI-Technologieakzeptanzmodell
XAIT	Explainable AI Trust	Erklärbares KI-Vertrauen
XAI	Explainable Artificial Intelligence	Erklärbare Künstliche Intelligenz
PUF	Perceived Usefulness	Wahrgenommene Nützlichkeit
EOU	Ease of Use	Einfachheit der Nutzung
BI	Behavioral Intention	Verhaltensintention
CI	Collaborative Intention	Kollaborationsintention
FAM-TEC	Familiarity with Technology	Vertrautheit mit Technologie
CLT	Construal Level Theory	Konstruktebenen-Theorie
SEM	Structural Equation Model	Strukturgleichungsmodell
UV	Independent Variable	Unabhängige Variable
N	Sample Size	Stichprobengrösse

11 Anhang

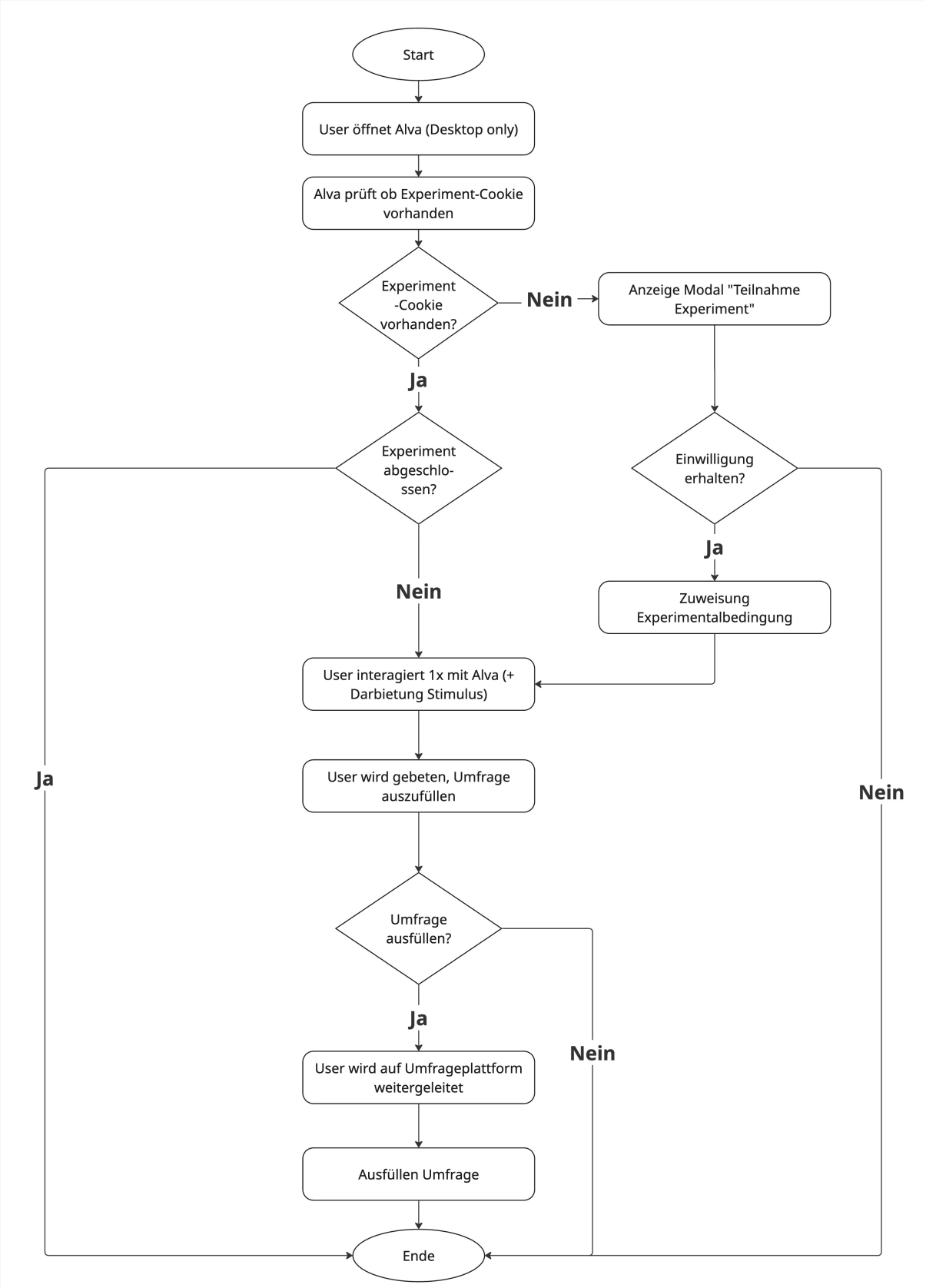

Abbildung 4

Ablauf des Experiments in drei Phasen