INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

FLLM 2025

ADRIAN RYSER, FLORIAN ALLWEIN, TIM SCHLIPPE

CALIBRATED TRUST IN DEALING WITH LLM HALLUCINATIONS: A QUALITATIVE STUDY

Vienna, AT

Nov 27, 2025

WHO WE ARE

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Adrian Ryser

— Student of Digital Transformation

— Baar (CH)

Florian Allwein

— Professor for Digital Transformation

— Berlin

Tim Schlippe

— Professor for Artificial Intelligence

— Karlsruhe

01

INTRODUCTION

DO WE TRUST LLMs TOO MUCH?

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- LLMs increasingly widespread and popular
- But they produce **hallucinations** – outputs that are factually incorrect yet appear plausible
- Users still trust LLM output – only 27% of them verify it (Alich 2025)

- **RQ: How do experiences of LLM hallucinations influence users' trust in LLMs and users’ interaction with LLMs?**

D. Alich, “EY AI Sentiment Index 2025,” Apr. 2025. [Online]. Available: https://www.ey.com/de_de/functional/forms/download/2025/04/ey-ai-sentiment-index-2025

02

RELATED WORK

PREVIOUS RESEARCH

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

LLM hallucinations

- research on **mitigating** hallucinations: e.g. Huang et al. (2025), Tonmoy et al. (2024)
- hallucinations cannot be fully eliminated, as no model can produce factually correct outputs for all possible inputs (Xu et al. 2025)
  - → Hallucinations as **system-inherent properties** of LLMs
- users need to **trust** tools to some extent

Trust in LLMs

- Previous studies: factors like **source references** or **text that sounds convincing** influence users‘ trust
- typically conducted in controlled experiments
- users’ personal experiences, individual strategies, and the everyday use of LLMs not examined

→ we need to better understand user behavior and verification strategies in response to hallucinations

Z. Xu, S. Jain, and M. Kankanhalli, “Hallucination is Inevitable: An Innate Limitation of Large Language Models,” Feb. 13, 2025, *arXiv*: arXiv:2401.11817. doi: 10.48550/arXiv.2401.11817.
L. Huang et al., “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.
S. M. T. I. Tonmoy et al., “A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models,” Jan. 08, 2024, arXiv: arXiv:2401.01313. doi: 10.48550/arXiv.2401.01313.

OUR MODEL

- Lee & See 2004:
  - **Undertrust** (e.g. missed opportunities)
  - **overtrust** (e.g. uncritical acceptance)
  - **calibrated trust** (alignment user reliance <-> system performance)
- Blöbaum 2022: Trust as a **dynamic process**, not measurable quantity
- Afroogh et al. 2024: trust factors in human-Machine interaction: **technical/ human/ contextual/ axiological** (values)

Combined here:

- **recursive** trust calibration process
- past interactions shape future trust judgments, especially in uncertain/ unfamiliar situations.

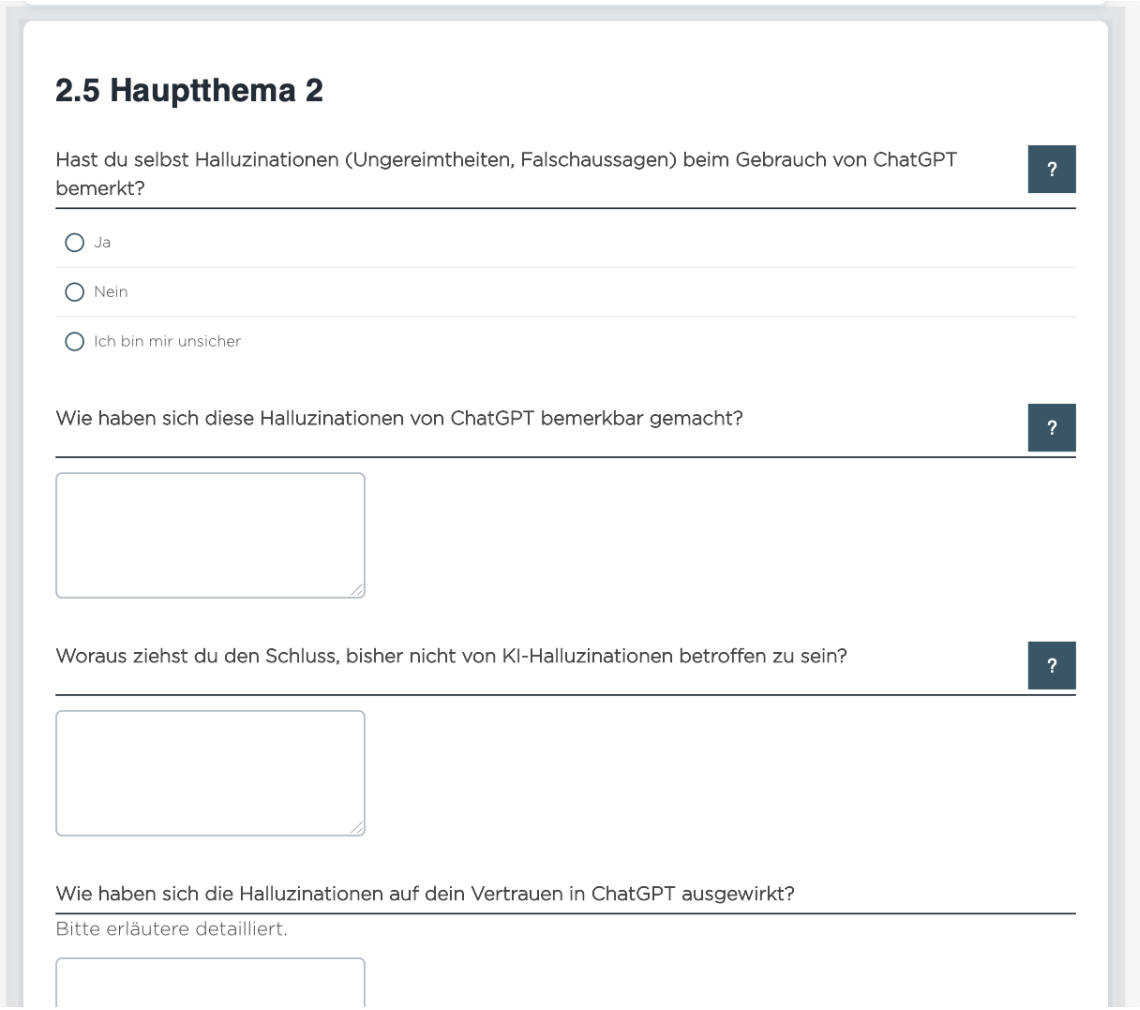
- S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi, “Trust in AI: progress, challenges, and future directions,” Humanit. Soc. Sci. Commun., vol. 11, no. 1, p. 1568, Nov. 2024, doi:10.1057/s41599-024-04044-8
- B. Blöbaum, Vertrauen, Misstrauen und Medien. Wiesbaden: Springer Fachmedien Wiesbaden, 2022. doi:10.1007/978-3-658-38558-3
- J. D. Lee and K. A. See, “Trust in Automation: Designing for Appropriate Reliance,” Hum. Factors, vol. 46, no. 1, pp. 50–80, Mar. 2004, doi: 10.1518/hfes.46.1.50_30392

03

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

SURVEY

SURVEY



2.5 Hauptthema 2

Hast du selbst Halluzinationen (Ungereimtheiten, Falschaussagen) beim Gebrauch von ChatGPT bemerkt?

- Ja
- Nein
- Ich bin mir unsicher

Wie haben sich diese Halluzinationen von ChatGPT bemerkbar gemacht?

Woraus ziehst du den Schluss, bisher nicht von KI-Halluzinationen betroffen zu sein?

Wie haben sich die Halluzinationen auf dein Vertrauen in ChatGPT ausgewirkt?

Bitte erläutere detailliert.

RQ: **How do experiences of LLM hallucinations influence users' trust in LLMs and users' interaction with LLMs?**

- Qualitative survey
- Promoted via LinkedIn/ IU networks
- Participants could respond at their own time and in their usual surroundings.
- **192 responses** 🌟

QUALITATIVE DATA ANALYSIS FOLLOWING MAYRING

Kode	Anz.	Kommentar
○ Kategoriessystem	886	Eigenes Kategoriensystem welches durch deduktives Ableiten aus der Theorie und induktiver Erzeugung von Kategorien und Codes direkt aus den Daten aufgebaut ist.
○ Vertrauen	141	Die Hauptkategorie Vertrauen umfasst die Aussagen mit Relevanz zum Konzept des Vertrauens. Dabei sind auch Vertrauensveränderung von Interesse.
○ Grundsätzliches Vertrauen	23	Entstehung Induktiv. Definition: Probanden welche Antworten von ChatGPT grundsätzlich oder mehrheitlich vertrauen.
○ Misstrauen	37	Diese Kategorie fasst Gründe welche auf Misstrauen, Skepsis oder Ablehnung schließen lassen.
○ Datenschutz / Ethik	5	Entstehung Indikativ. Definition: Bedenken bezüglich des Datenschutzes der Inputdaten oder ethische Bedenken bezüglich der Nutzung von Lerndaten.

- Systematic coding process, supported by **Atlas.ti** software
- **Categories** developed iteratively
- Combination of hermeneutic circle & **Mayring**‘s (2023) qualitative content analysis

P. Mayring, Einführung in die qualitative Sozialforschung., 7. Aufl. Beltz Verlagsgruppe, 2023.

04

RESULTS

RESULTS: SEE OUR REPOSITORY

lfdn	dispcode	lastpage	duration	v_nutzun	v_nutzun	v_abo	v_info_art		v_hallu	v_halluzinationen_e
16	31	112691	351	1	-66	2		Quelle	1	Quellangaben
17	31	112691	83	2	-66	2	Allerlei; Alltagsfragen; Spez. Fragen fürs Studi		3	-66
18	31	112691	601	2	-66	2		, man	1	Siehe vorherige Ant
19	31	112691	945	2	-66	2		die eir	1	Steht in der vorherig
20	31	112691	267	3	-66	2	-99		2	-66
21	31	112691	439	1	-66	1		ntniss	1	Quellangaben, ode
22	31	112691	225	2	-66	2	Beim Erstellen von Hausarbeiten (Gliederung e		1	Antwort passte nich
23	31	112691	410	1	-66	1		selbst	1	Literaturvorschläge
24	31	112691	635	2	-66	2	Ich benutze es hauptsächlich wenn ich allgeme	ausart	1	Einmal habe ich ihm
26	31	112691	224	1	-66	1			1	Unrichtige aussagen
27	31	112691	185	2	-66	1	Sätze umformulieren (Studium); Ideen für Sem		1	3 Arme in Bilder
28	31	112691	1150	2	-66	1		auch e	1	Immer gleiche antw
29	31	112691	835	2	-66	2	Korrespondenz, Skript, Ideen		3	-66
30	31	112691	305	2	-66	2			3	-66
31	31	112691	491	2	-66	2			2	-66
32	31	112691	260	1	-66	1	Hausarbeit, alltäglicher Rat etc.	ebniss	1	Bei Mathematik ode
33	31	112691	577	1	-66	1		iziert	1	Z.B. wenn ich einen
34	31	112691	566	2	-66	2	Für fachlichen Austausch im Studium, Dialoge		3	-66
37	31	112691	283	1	-66	2		rüfung	1	IU Wissenstest
38	31	112691	781	2	-66	1	Um quellen für meine Hausarbeit zu finden Au	Softw	1	- durch Widersprüch
39	31	112691	1135	1	-66	1		ige M	1	Wenn man Codings
40	31	112691	300	2	-66	2	Schreibst von textvorschlägen für die masterth	ma K	1	Bei literaturrecherch
41	31	112691	5533	2	-66	2		tionsq	1	Durch Nutzung and
43	31	112691	889	1	-66	1	Bilder für Homepage erstellen, Briefe/Antworte	hmal	1	In dem ChatGPT ur
44	31	112691	848	2	-66	1		chartik	3	-66
45	31	112691	311	2	-66	2	Ich nutze es als Unterstützung für Social Media		3	-66
46	31	112691	328	2	-66	2		en ge	1	Falsche Quellenang
47	31	112691	272	1	-66	2			1	Ich habe Fotos von
48	31	112691	1434	1	-66	2	Ideen sammeln, übersetzen, Zusammenfassur	rt. Akt	1	ChatGPT kann bsp
49	31	112691	233	1	-66	1			1	Angabe falscher Lit
50	31	112691	1018	2	-66	1	Ideensammlung für Seminararbeit, Vorschläge	ortet (	1	Siehe vorhin. Falsch
51	31	112691	3273	1	-66	1		ß. Viel	1	Teilweise wusste ich
52	31	112691	146	3	-66	2	Inspiration und Gedankenanstosse für Arbeiter		2	-66
53	31	112691	399	2	-66	2			1	Personen im Film m
54	31	112691	809	2	-66	1	Arbeit, Studium, anstatt Google	edien	1	Manchmal hat er Au
55	31	112691	833	2	-66	2		n KI f	3	-66
56	31	112691	552	2	-66	2	Mittlerweile als Ersatz für Google. Aber auch fü	PT an	1	Die Links waren nicl
57	31	112691	1026	1	-66	1		r unte	1	- Im Studium gibt es
59	31	112691	545	2	-66	1	Ich nutze ChatGPT für die Formulierung von T		1	unbrauchbare Antw
							Hauptsächlich für Hausarbeiten oder generelle			

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

doi.org/10.5281/zenodo.15618622

RESULTS

- 82% familiar with LLM hallucinations
- 68% reported personal experience
- Hallucination experiences did **not lead to a general loss of trust**.
- Changes in how users interact with LLMs

"ChatGPT always makes

"It made me a bit

"If it's ok to be wrong, then I trust the information 100%, but if it has to be correct, then it's more like 70-80%". (A17)

RESULTS

- Trust was not lost, but **recalibrated** based on *prior experience* and the perceived relevance of the task.

“Little trust, especially with important topics” (A64)

”Damn, now I have even less trust. Especially because I had a lot of texts summarized. Let's see if I'll stop using ChatGPT altogether.” (A99)

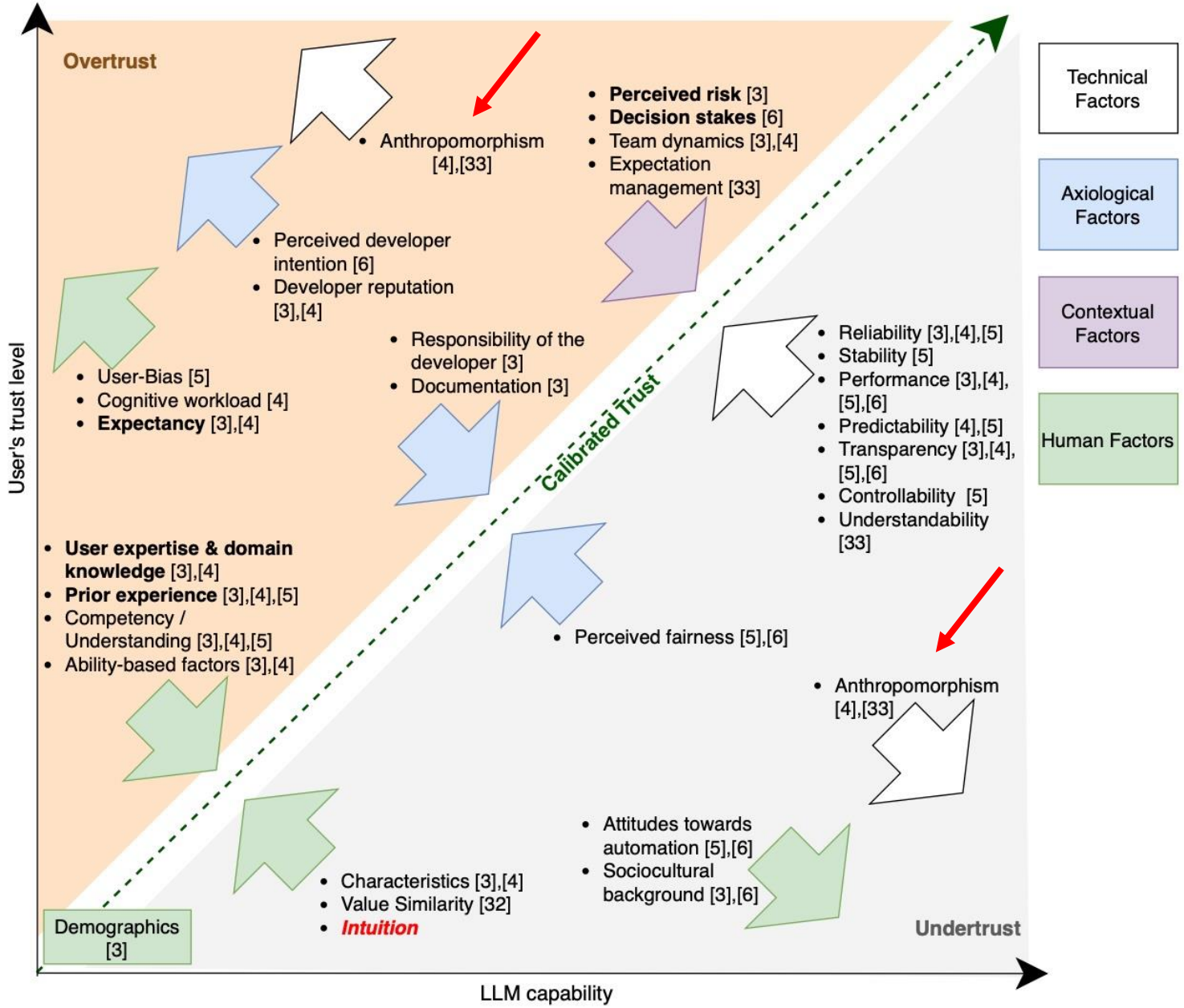
RESULTS

- **Participants adapt their trust based on *prior experience*, *perceived risk*, and *decision stakes***
- In contexts like academic work, where accuracy is critical: limited use of LLMs

"[I use ChatGPT] less as a source of information, I switched back to normal 'googling'" (A9)

"The greater the impact of an incorrect answer, the less I can check the correctness of the answer using my own knowledge and the more illogical the answers seem to me, the more likely I am to verify the answers." (A32)

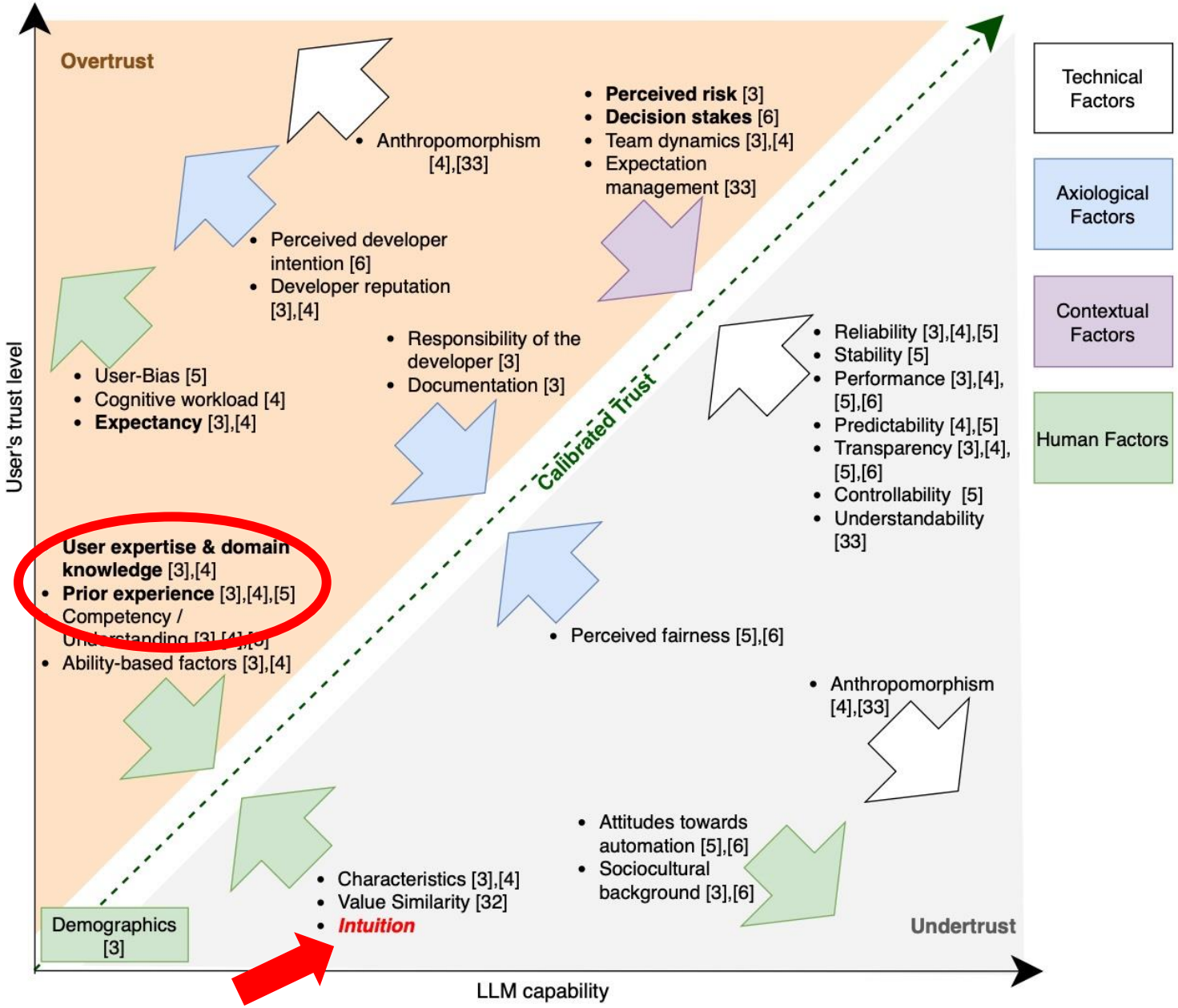
RESULTS: CALIBRATED TRUST



calibrated trust (appropriate level of trust):

- trust calibration e.g. influenced by **user expertise & domain knowledge**, prior experience with LLMs
- Some factors debated or context-dependent: e.g. **anthropomorphism** may foster overtrust or undertrust depending on user expectations

RESULTS: CALIBRATED TRUST

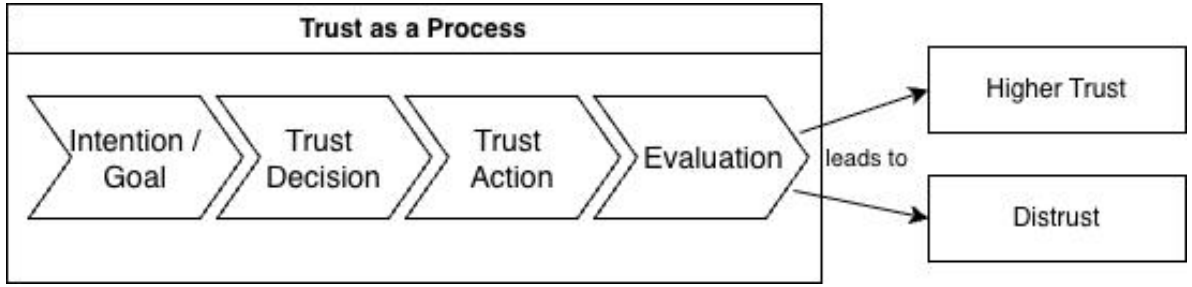


- We confirmed e.g.
  - prior experience
  - user expertise & domain knowledge

  … as user-related trust factors
- & found **intuition** as an additional factor

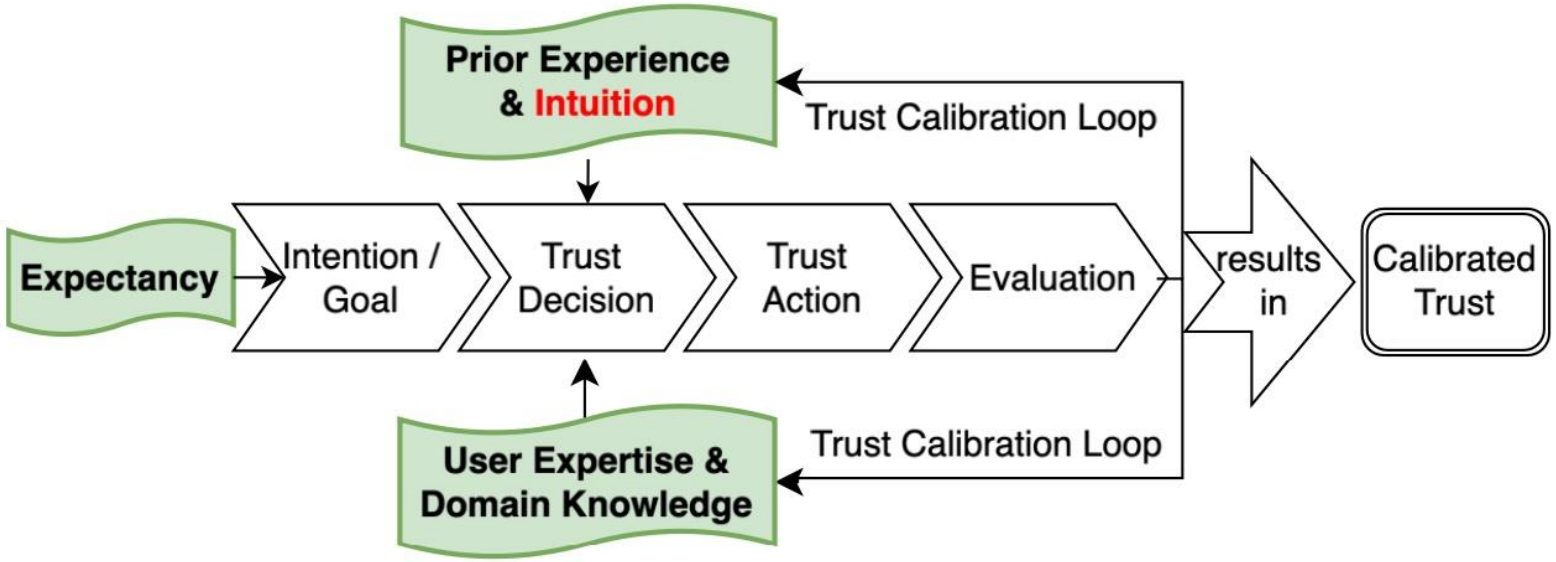
RESULTS: RECURSIVE TRUST CALIBRATION

- We adapted and expanded Blöbaum's recursive trust calibration process to the case of hallucination-prone LLMs.



Blöbaum 2022

RESULTS: RECURSIVE TRUST CALIBRATION



- Trust factors evolve through repeated user interaction with LLMs
- Repeated calibration loops lead to **calibrated trust**, also: increased **AI literacy**

05

CONCLUSION/ FUTURE WORK

CONTRIBUTIONS

- how hallucinations in LLMs affect users' trust and which strategies they use to adjust, stabilize, or recalibrate it.
- While outputs that are professionally or personally significant are often verified, responses to everyday questions or minor tasks are typically accepted without further verification.
- Trust in LLMs is not a fixed state, but a dynamic, experience-based process.
- *Intuition* supports this process, particularly in situations where LLM output verification is not feasible or when *perceived risk* of the current task is low.

Recommendations for users

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

01

CALIBRATE TRUST

Actively calibrate trust considering the task's relevance and user's level of domain knowledge

02

VERIFY CONTEXTUALLY

Tailor verification efforts to perceived risk and relevance of the task

03

INTEGRATE INTUITION

Based on prior experience, rely on intuition to detect hallucinations

04

BUILD AI LITERACY

Develop better understanding of how LLMs function and where their limitations lie

05

TREAT LLMS AS ASSISTANTS

See LLMs as a supporting tool, not primary source of information

FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- Behavioral sources such as chat logs to deepen understanding of LLM use
- Longitudinal studies to show how trust evolves over time
- Emotional responses such as frustration or irony

- Limited awareness of hallucinations among users of LLMs
- Hallucinations are an inherent property of LLMs

→ **we call on researchers and users of AI and LLMs to be mindful of them and adopt their use accordingly**

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Q & A

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

THANK YOU

Florian Allwein

Professor Digital Transformation at IU International University of Applied Scienc...

IU Internationale Hochschule GmbH
Juri-Gagarin-Ring 152
D-99084 Erfurt

Florian Allwein
ask me
florian.allwein@iu.org

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

doi.org/10.5281/zenodo.15618622

BONUS: HALLUCINATIONS PERSIST (OPENAI)

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate

arXiv > cs > arXiv:2509.04664

Search... All fields Search

Help | Advanced Search

Computer Science > Computation and Language

[Submitted on 4 Sep 2025]

Why Language Models Hallucinate

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious -- they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded -- language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:2509.04664 **[cs.CL]**
(or arXiv:2509.04664v1 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2509.04664

Submission history

From: Adam Kalai [view email]
[v1] Thu, 4 Sep 2025 21:26:31 UTC (142 KB)

Access Paper:
View PDF
HTML (experimental)
TeX Source
Other Formats

Current browse context:
cs.CL
< prev | next >
new | recent | 2025-09
Change to browse by:
cs

References & Citations
NASA ADS
Google Scholar
Semantic Scholar

Export BibTeX Citation

Bookmark

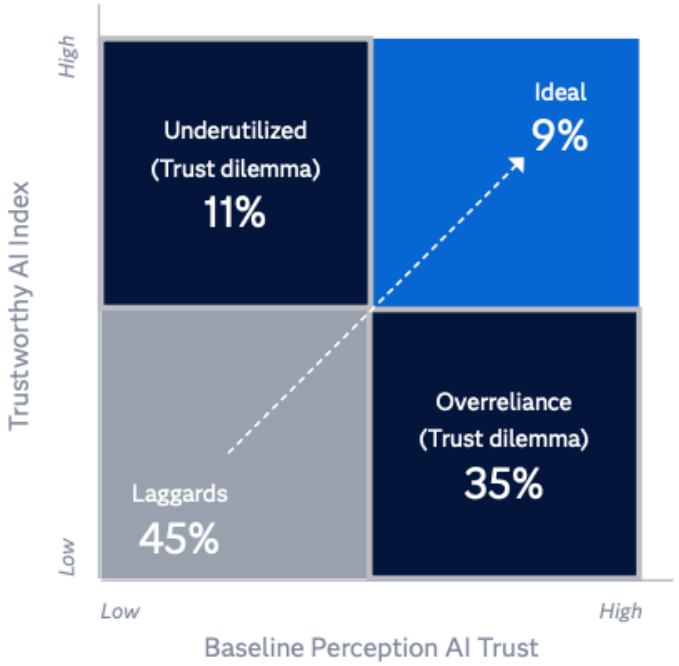
- https://arxiv.org/abs/2509.04664
- https://openai.com/index/why-language-models-hallucinate/

BONUS: THE TRUST DILEMMA (SAS)

The chart below plots the relationship between the perceived trust in AI systems and their actual trustworthiness, illustrating the "trust dilemma." This misalignment, evident across all regions, represents a critical barrier to effective AI adoption. Most organizations experience this misalignment, with relatively few achieving the ideal balance. Two risks emerge: underutilization of reliable systems when trust remains low and overreliance on unproven systems when confidence is disproportionately high. The challenge is particularly acute for generative AI, where rapid enthusiasm has outpaced governance and data quality.

GLOBAL TRUST DILEMMA

The matrix presents clear categories, but both trust in AI and its trustworthiness lie on a continuum. While the report uses a 2x2 framework, readers should keep in mind that shifts between levels are gradual, not binary.



– https://www.sas.com/de_at/news/press-releases/2025/september/ai-trust-idc-study.html