

# Speaking the Right Language: The Impact of Expertise (Mis)Alignment in User-AI Interactions

Shramay Palta<sup>†</sup>◇\* Nirupama Chandrasekaran<sup>◇</sup>  
Rachel Rudinger<sup>†</sup> Scott Counts<sup>♣</sup>

<sup>†</sup>University of Maryland, College Park      ◇Microsoft Research  
♣Microsoft Corporation  
{spalta,rudinger}@cs.umd.edu  
{niruc,counts}@microsoft.com

## Abstract

Using a sample of 25,000 Bing Copilot conversations, we study how the agent responds to users of varying levels of domain expertise and the resulting impact on user experience along multiple dimensions. Our findings show that across a variety of topical domains, the agent largely responds at proficient or expert levels of expertise (77% of conversations) which correlates with positive user experience regardless of the user’s level of expertise. Misalignment, such that the agent responds at a level of expertise below that of the user, has a negative impact on overall user experience, with the impact more profound for more complex tasks. We also show that users engage more, as measured by the number of words in the conversation, when the agent responds at a level of expertise commensurate with that of the user. Our findings underscore the importance of alignment between users and AI when designing human-centered AI systems, to ensure satisfactory and productive interactions.

## 1 Introduction

We have seen significant advancements in LLM development, which has enhanced the capabilities of model-based agents like ChatGPT that allow them to excel on tasks ranging from quick information retrieval to more creative or technical pursuits such as drafting essays, writing code, and designing artworks. These capabilities are not only attested to by their performance on standardized benchmarks but are also reflected in their use across a diverse set of real-world domains (Suri et al., 2024). Recently developed LLMS are able to assist humans across a variety of fields like teaching (Wang et al., 2024; Alsafari et al., 2024) and the clinical domain (Han et al., 2024), and have led to an increase in user productivity (Peng et al., 2023; Cambon et al., 2023).

\*Work done during an internship at Microsoft Research.

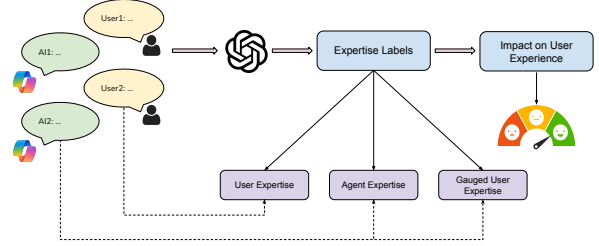


Figure 1: An overview of our expertise classifier pipeline.

However, while millions of users utilize these models for a variety of tasks, their expectations, backgrounds, and interactions with these tools can differ significantly. One key aspect where users might differ is their domain expertise in the conversation topic with the agent LLM.<sup>1</sup> Not all end users share the same level of domain knowledge and thus may have different preferences and abilities to process the information that the model would return to them. A “beginner” user might want simple, general purpose information on a topic and could possibly be overwhelmed if presented with too “high-level” information. On the other hand, a domain “expert” might have an unsatisfactory experience with the LLM if not given a deeper and detailed response.

Thus, we ask: *What is the ideal expertise level of the LLM, and what are the consequences of any misalignment between the user and the LLM on domain expertise?*

To answer this, we develop an ordinal 5-point scale-based expertise classifier (shown in Figure 1) that we apply to a corpus of over 25,000 Bing Copilot conversations sampled across a variety of domains. We generate three measures of expertise for each conversation. First and second, we classify the level of expertise of the user and the LLM respectively in the topic of the conversation.

<sup>1</sup>Here on called LLM.

Domain	Count
Creative Writing and Editing	6761
Programming and Scripting	6078
Business and Finance	4090
Professional Writing and Editing	2269
Entertainment	1406
Mathematics and Logic	1306
Data Analysis and Visualization	1245
Biology	994
Sports and Fitness	484
Machine Learning and AI	400
<b>Total</b>	<b>25033</b>

Table 1: Domain-wise distribution of Copilot conversations used in our study.

Third, we classify the gauged expertise of the user, defined as the judged level of expertise of the user based on the responses made by the LLM. We show similarities and differences in these three types of expertise labels within the same Copilot conversation to identify cases where the LLM is aligned or misaligned with the user. We then assess the impact of (mis)alignment on three measures of user interaction experience: user satisfaction, level of engagement, and complexity of task.

## 2 Methodology

Consider a conversation  $C_i$  from a corpus of conversations  $C$  consisting of  $t$  interaction turns of user-agent utterances  $C_i = [U_1, A_1, \dots, U_t, A_t]$ .<sup>2</sup> We take a random sample of such conversations from Bing Copilot from the month of June 2024 with at least  $t \geq 2$  interaction turns for both the user and the agent, yielding a set of 677,801 conversations. Using the preprocessing steps discussed in Appendix A.1, we get a final set of 25,033 conversations. We present the distribution of the conversations across different domains in Table 1.

### 2.1 Expertise Labels

We expect that users of varying expertise levels interact with Bing Copilot, and thus in order to determine the (mis)alignment between the user and the LLM expertise, we compute three different types of expertise labels as follows:

**User Expertise:** Expertise of the user in the conversation domain based on the User ( $U_t$ ) only side of a conversation  $C_i$ .

**Gauged User Expertise:** Predicted expertise of the user in the conversation domain, based on the

<sup>2</sup>Here  $U_t$  refers to a *user* utterance and  $A_t$  refers to an *AI Agent* utterance.

Domain	Mean SAT Score
Creative Writing and Editing	3.03
Programming and Scripting	14.42
Business and Finance	17.22
Professional Writing and Editing	19.55
Entertainment	11.93
Mathematics and Logic	-11.03
Data Analysis and Visualization	13.35
Biology	18.42
Sports and Fitness	16.24
Machine Learning and AI	16.59

Table 2: Mean SAT Scores for conversations across different domains.

LLM ( $A_t$ ) only side of a conversation  $C_i$ .

**Agent Expertise:** Expertise of the agent in the conversation domain, based on the LLM ( $A_t$ ) only side of a conversation  $C_i$ .

We prompt GPT-4-Turbo (Achiam et al., 2023) to compute these three types of expertise labels for each conversation using a 5-point ordinal scale as “**Novice**”, “**Beginner**”, “**Intermediate**”, “**Proficient**”, and “**Expert**”.<sup>3</sup> We present the definitions of each of these labels along with the human-validated system prompts in Figure 9, Figure 10 and Figure 11. Details on the human validation of the predicted expertise labels can be found in Appendix A.5.

### 2.2 Metrics for User Experience

We use the following three metrics to understand the impact of expertise (mis)alignment on the user interaction experience:

- **SAT Score:** Lin et al. (2024) introduced SPUR (Supervised Prompting for User satisfaction Rubrics), an iterative prompting framework using supervision from labeled examples to estimate the user satisfaction score from a multi-turn conversation with an LLM agent. SPUR enables the computation of both satisfaction and dissatisfaction scores. We adopt this framework and define the overall satisfaction score, denoted as SAT, as the difference between the satisfaction and dissatisfaction scores. The SAT score ranges from -100 to 100. The SAT score rubric is human-validated, the details of which are mentioned in Lin et al. (2024). We present the mean SAT scores for the different domains in our conversation corpus in Table 2.

<sup>3</sup>We refer the readers to § 6 for more insights into why a single model was chosen for computing the labels.

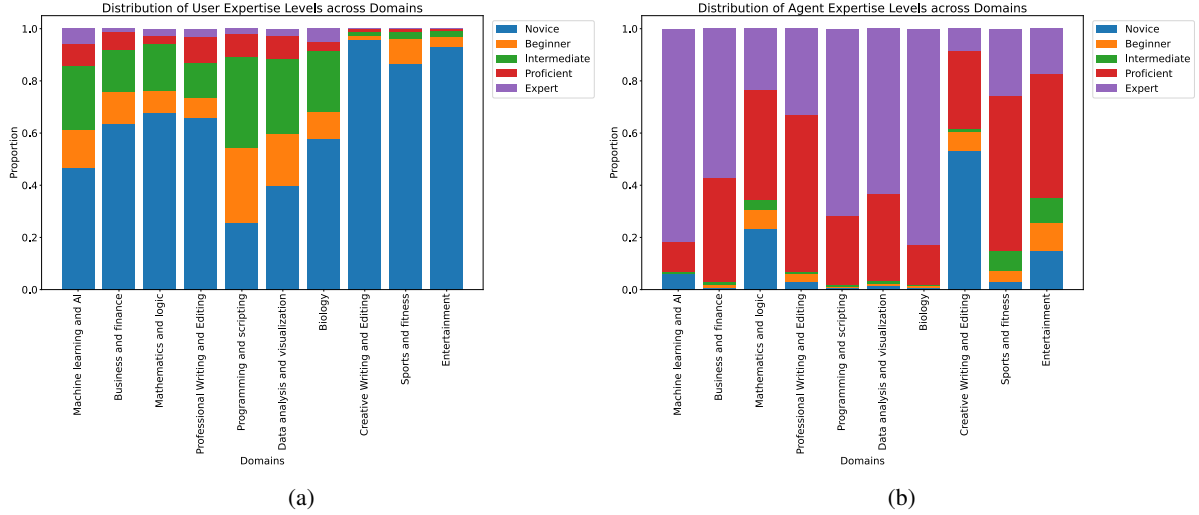


Figure 2: Barplots showing the distribution of User Expertise (left) and Agent Expertise (right) on different domains of Copilot conversations.

- **Task Complexity:** [Suri et al. \(2024\)](#) introduced a task complexity classification method based on Anderson and Krathwohl’s Taxonomy of learning domains ([Armstrong, 2010](#)) which categorizes the task complexity into six levels from lowest complexity to highest: *Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*. For simplicity, we group *Remember* and *Understand* as Low Complexity and *Apply*, *Analyze*, *Evaluate* and *Create* as High Complexity tasks. The Task Complexity metric is human-validated, the details of which are mentioned in [Suri et al. \(2024\)](#).
- **Conversation Length:** As all our conversations are multi-turn, we look at the number of words across all user turns as a proxy for the user engagement level.

### 3 User and LLM Expertise

Using the labels described in § 2.1, we compute the user expertise, the agent expertise, and the gauged user expertise on our set of 25, 033 conversations and present the distributions of these expertise labels across different domains in [Figure 2a](#), [Figure 2b](#) and [Figure 5](#) respectively.<sup>4</sup>

Overall, we observe that a majority of the users (63.9%) are labeled as “Novice” on our ordinal scale, with a small number of users being classified as “Proficient” (5.2%) or “Expert”(1.6%). These small numbers of “Proficient” and “Expert” users occur in the more technical domains (like Programming and scripting) as compared to the non-

technical domains (like Entertainment). We also see that, in a majority of the cases, the LLM gets labeled as “Proficient”(34.9%) or “Expert”(42.4%). Once again, we observe the number of “Proficient” and “Expert” labels to be fewer for non-technical domains as compared to the technical domains. Overall, [Figure 6](#) shows that the user has a lower expertise than the LLM in a majority (80.1%) of the conversations. This makes sense in that most users are likely to be non-experts, while the model should have higher expertise in order to provide value to the user.

Finally, for the gauged user expertise ([Figure 5](#)), we observe that a majority of the users are labeled as “Intermediate”(37.2%) or above. Comparing this finding with the User Expertise label distribution ([Figure 6](#)) indicates that there are cases of user overestimation, in which user expertise as gauged by the response from the LLM is higher than when directly assessing user expertise (57.4% of conversations), and user underestimation, in which gauged expertise is lower than the user expertise (4.23% of conversations). We also test our expertise classifier on a sample of WildChat ([Zhao et al., 2024](#)) conversations in [Appendix A.4](#), where we observe a similar distribution of labels, hence demonstrating the generalisability of our expertise classifier.

### 4 Impact on User Experience

To understand how expertise (mis)alignment impacts the user experience, we do the following:

First, we fit a piecewise regression of the dif-

<sup>4</sup>We refer the readers to [Table 3](#) for more details.

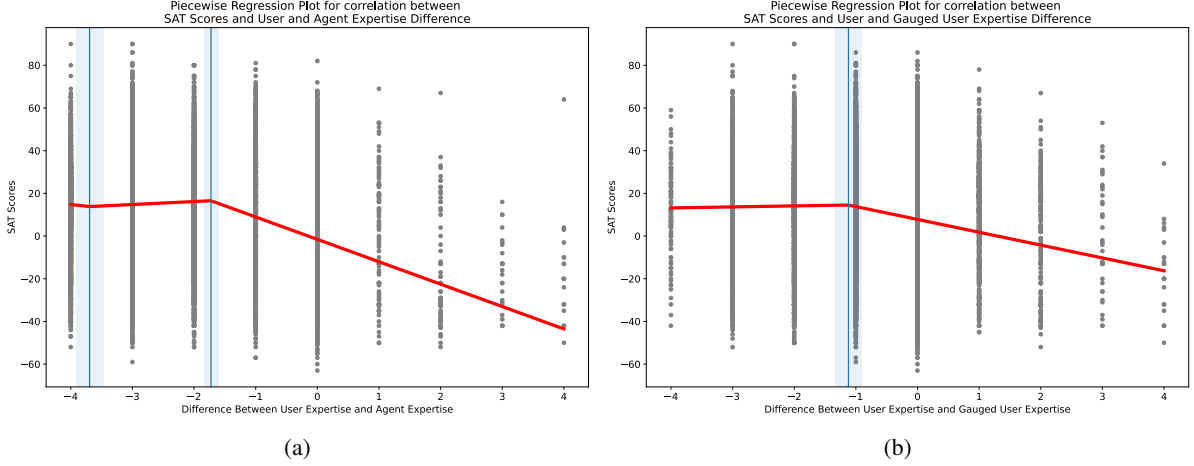


Figure 3: Piecewise regression plots showing the correlation between Expertise Difference and SAT scores.

ference between user and LLM expertise on user satisfaction. Figure 3a shows that when the user expertise exceeds the LLM expertise, there is a negative impact ( $R^2 = 0.11, p = 1.36E^{-141}$ ) on the SAT score, with the overall SAT score becoming negative in absolute terms once the user is one level more expert than the LLM. Increasing gaps in user expertise over LLM expertise are associated with continually lower user satisfaction scores. Figure 7 also highlights that it is preferable for the LLM to be quite expert regardless of the user’s level of expertise, with lower than “Proficient” responses from the LLM generally leading to decreased user SAT scores. Further, underestimating the user’s expertise level (Figure 3b), as measured by gauged user expertise, has a clear negative impact ( $R^2 = 0.03, p = 1.06E^{-26}$ ) on the SAT score, also highlighted in Figure 7.

The nature of the user’s task may impact the effect of expertise (mis)alignment, and we operationalize one aspect of this with the task complexity measure. We fit linear regressions of the gap in user to LLM expertise on user satisfaction separately for conversations with low versus high complexity tasks, with high and low task complexity defined as in § 2.2. As seen in Figure 4a, for low complexity tasks, the user-LLM expertise gap had essentially no effect on user satisfaction ( $r = -0.04$ ), while for high complexity tasks, this effect was of medium effect size ( $r = -0.29$ ). The effect of user expertise underestimation as reflected in the gap between actual and gauged user expertise (Figure 4b) shows a similar if somewhat moderated trend with no effect ( $r = 0.02$ ) for tasks of low complexity and a small effect for tasks of high

complexity ( $r = -0.15$ ).

Finally, turning to the impact of user-LLM (mis)alignment on user engagement, Figure 8 shows that the amount of engagement, as defined in § 2.2, increases along with the level of user expertise. That is, more expert users tend to have longer conversations. This effect appears to depend on the level of expertise of the LLM, however, such that users at each level of expertise tend to engage more when the LLM responds at a similar level of expertise. “Proficient” and “Expert” users tend to engage relatively more with a Proficient or Expert LLM, while “Novice” and “Beginner” users tend to engage more with a Novice or Beginner LLM. Thus, while the user satisfaction measure suggests that users prefer a more expert LLM, they engage more with LLMs of commensurate expertise.

## 5 Related Works

The term “expertise” has been defined along multiple dimensions (Bourne Jr et al., 2014; Garrett et al., 2009), such as the “extent and organization of knowledge and special reasoning processes to development and intelligence” (Feltovich and Hoffman, 1997). Traditionally, expertise has always been correlated with knowledge, skill and other cognitive concepts (Bourne Jr et al., 2014). Building upon this idea, Desmarais et al. (1995) introduced a probabilistic approach to model user expertise, and Ferrod et al. (2021) did so in a dialogue based setting. Different methods like heuristic rules (Vaubel and Gettys, 1990) have also shown promise in inferring expertise for word processing tasks. Notably, we model the user, agent and the gauged user expertise from conversational data to show

the impacts of expertise mis-alignment on the user interaction experience.

Additionally, many works have explored how LLMs are mis-aligned with humans across an axis of different dimensions like moral judgements (Hendrycks et al., 2023; Jiang et al., 2022), cultural and societal norms (Palta and Rudinger, 2023; Acquaye et al., 2024; Naous et al., 2024; Bhatia and Shwartz, 2023; Huang and Yang, 2023), healthcare (Levy et al., 2024) and notions of plausibility (Palta et al., 2024, 2025). Similarly, we show that LLMs are misaligned with humans along notions of expertise, which can lead to unsatisfactory user experiences.

## 6 Conclusion

We examine the alignment between LLMs and users along a dimension relevant to the user experience: expertise. We show that the LLM’s expertise is largely proficient or expert, which correlates with positive user satisfaction and exceeds user expertise in a majority of the cases. Further, underestimating the user’s level of expertise correlates with lower and even negative user satisfaction, with the effect stronger for more complex tasks. Users tend to engage more, however, when the LLM responded at a level of expertise similar to their own, suggesting that the system strike a balance between generally high expertise which is liked by all users and matched expertise to best engage users. Future work may explore intervention strategies to strike this balance and mitigate obvious cases of user underestimation in real time.

## Limitations

Our analysis of user and LLM expertise misalignment and its downstream impacts is based on predicted expertise labels and predicted user satisfaction scores (SAT). While we human validated the classification labels, there is still a possibility of some errors which could impact the results. We study conversations only in English, which may limit the generalizability our findings, given that a lot of conversations with Copilot take place in non-English languages. Our analysis is limited to English conversations to facilitate human-validation of predicted expertise labels. Future work may consider extending our findings in the multi-lingual domain. We use the same prompts to predict user and agent expertise on all the topical domains of our conversations. While more personalized tem-

plates might be able to capture the notions of expertise more accurately for different domains, we use the same template throughout to be able to make a fair comparison across all our experimental settings. Additionally, we only use GPT-4 to predict the expertise labels. While it is possible that other LLMs might be able to judge the expertise better, we restrict ourselves to the same model as the SAT rubric and the task complexity classifier to maintain model consistency. Future work may consider extending our expertise classifier to evaluate the alignment between humans and multiple different LLMs. Finally, all our results are correlational, but do indicate that a mismatch in expertise between the user and LLMs could be one of the causes of user dissatisfaction. We hope our findings motivate future works to involve experiments where agent expertise can be manipulated to determine whether it has a causal impact on user satisfaction or not.

## Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback on this paper. We would also like to thank Mengting Wan, Longqi Yang, David Tittsworth, Sujay Kumar Jauhar, Bahar Sarrafzadeh, Sahithya Ravi, Mohit Chandra, Omar Shaikh, Anubhav Jangra, Rupak Sarkar and Ishani Mondal for their valuable comments and suggestions.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. [Susu box or piggy bank: Assessing cultural commonsense knowledge between Ghana and the US](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9483–9502, Miami, Florida, USA. Association for Computational Linguistics.
- Bashaer Alsafari, Eric Atwell, Aisha Walker, and Martin Callaghan. 2024. [Towards effective teaching assistants: From intent-based chatbots to llm-powered teaching assistants](#). *Natural Language Processing Journal*, 8:100101.
- Patricia Armstrong. 2010. Bloom’s taxonomy. *Vanderbilt University Center for Teaching*, pages 1–3.
- Mehar Bhatia and Vered Shwartz. 2023. [GD-COMET: A geo-diverse commonsense inference model](#). In

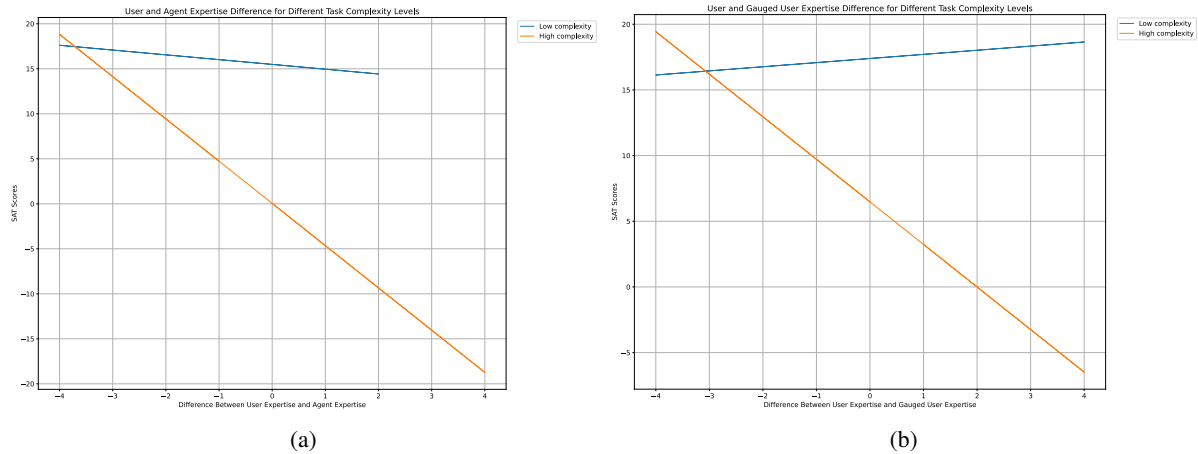


Figure 4: Regression plots showing the correlation between Expertise Difference and SAT scores for low and high complexity tasks.

*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001, Singapore. Association for Computational Linguistics.

Lyle E Bourne Jr, James A Kole, and Alice F Healy. 2014. Expertise: defined, described, explained.

Alexia Cambon, Brent Hecht, Ben Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, et al. 2023. Early llm-based tools for enterprise information workers likely provide meaningful boosts to productivity. *Microsoft Research. MSR-TR-2023-43*.

Michel C Desmarais, Ameen Maluf, and Jiming Liu. 1995. User-expertise modeling with empirically derived probabilistic implication networks. *User modeling and user-adapted interaction*, 5:283–315.

Paul J Feltoovich and Robert R Hoffman. 1997. *Expertise in context*. AAAI Press Menlo Park, CA.

Roger Ferrod, Federica Cena, Luigi Di Caro, Dario Mana, and Rossana Grazia Simeoni. 2021. [Identifying users’ domain expertise from dialogues](#). In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP ’21*, page 29–34, New York, NY, USA. Association for Computing Machinery.

SK Garrett, Barrett S Caldwell, EC Harris, and MC Gonzalez. 2009. Six dimensions of expertise: A more comprehensive definition of cognitive expertise for team coordination. *Theoretical Issues in Ergonomics Science*, 10(2):93–105.

Jiyeon Han, Jimin Park, Jinyoung Huh, Uran Oh, Jaeyoung Do, and Daehye Kim. 2024. Ascleai: A llm-based clinical note management system for enhancing clinician productivity. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. [Aligning ai with shared human values](#). *Preprint*, arXiv:2008.02275.

Jing Huang and Diyi Yang. 2023. [Culturally aware natural language inference](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. [Can machines learn morality? the delphi experiment](#). *Preprint*, arXiv:2110.07574.

Sharon Levy, Tahilin Sanchez Karver, William Adler, Michelle R Kaufman, and Mark Dredze. 2024. [Evaluating biases in context-dependent sexual and reproductive health questions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5801–5812, Miami, Florida, USA. Association for Computational Linguistics.

Ying-Chun Lin, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. 2024. [Interpretable user satisfaction estimation for conversational systems with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11115, Bangkok, Thailand. Association for Computational Linguistics.

Shuyo Nakatani. 2010. [Language detection library for java](#).

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. [Having beer after prayer? measuring](#)

cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegrefe, Marine Carpuat, and Rachel Rudinger. 2024. [Plausibly problematic questions in multiple-choice benchmarks for commonsense reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3451–3473, Miami, Florida, USA. Association for Computational Linguistics.

Shramay Palta, Peter Rankel, Sarah Wiegrefe, and Rachel Rudinger. 2025. [Everything is plausible: Investigating the impact of llm rationales on human notions of plausibility](#). *Preprint*, arXiv:2510.08091.

Shramay Palta and Rachel Rudinger. 2023. [FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Demirer Mert. 2023. [The impact of ai on developer productivity: Evidence from github copilot](#). *Preprint*, arXiv:2302.06590.

Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. [The use of generative search engines for knowledge work and complex tasks](#). *Preprint*, arXiv:2404.04268.

Kent P Vaubel and Charles F Gettys. 1990. Inferring user expertise for adaptive interfaces. *Human-Computer Interaction*, 5(1):95–117.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. 2024. [Tutor copilot: A human-ai approach for scaling real-time expertise](#). *Preprint*, arXiv:2410.03017.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. [Wildchat: 1m chatgpt interaction logs in the wild](#). *Preprint*, arXiv:2405.01470.

## A Appendix

### A.1 Data Preprocessing

The requirement of two or more turns helped ensure we had sufficient signal to assess the level of expertise of both user and LLM. Additionally, the SAT score classifier (Lin et al., 2024) is also based on multi-turn conversations. We then use the `langdetect`<sup>5</sup> library (Nakatani, 2010) to determine the text language at each interaction turn

<sup>5</sup><https://pypi.org/project/langdetect/>

to further filter to conversations where the majority language detected overall was either English, or in case of a tie, English was one of the tied languages. The selection of English only conversations was done to ensure human judges could read a random sub-sample of conversations in order to human-validate our classifications of user and LLM expertise.

Using the domain classification methodology introduced in Suri et al. (2024), we further sampled the conversations at random from a set of topical domains to generate a final set of 25033 fully anonymized conversations. All personal, private, or sensitive information was scrubbed and masked before the conversations were used for this research. The access to the dataset is strictly limited to the authors who conducted hands-on analysis. This domain filtering step was performed to remove conversations in domains minimally or not at all relevant to the concept of expertise (e.g., travel).

**Ethics:** As part of the production process,<sup>6</sup> the Bing Copilot data is anonymized, and each conversation is formed by aggregating turns based on a unique conversation ID. Thus, none of the researchers who analyzed the data are able to recover and identify the conversations from any individual user. In addition, this research study was reviewed and approved by representatives from our institutional review board (IRB), as well as our ethics and security teams. No formal IRB certificate was required since we did not conduct any human studies for this work.

### A.2 Details on Computational Experiments

GPT-4 was run on a CPU machine and was allocated three hours to run all experiments with temperature 0. We did not perform a hyper-parameter search. All results are obtained from a single run. For detecting the conversation language, a V100 GPU was used for a total of 1 hour.

### A.3 Usage of AI Assistants

Our expertise classifier is based on GPT-4. We only use AI Assistants to assist our writing to identify grammar errors, typos and rephrase terms for readability.

<sup>6</sup>Our use of Copilot conversations falls within the Terms of Use outlined at <https://www.bing.com/new/termsofuse>

Label	Expertise		
	User	Gauged User	Agent
Novice	63.9%	25.7%	17.3%
Beginner	12.6%	14.7%	3.7%
Intermediate	16.7%	37.2%	1.7%
Proficient	5.2%	21.3%	34.9%
Expert	1.6%	1.1%	42.4%
Total	100%	100%	100%

Table 3: Percentage distribution of labels for User, Gauged User and LLM expertise on our set of 25033 Copilot conversations.

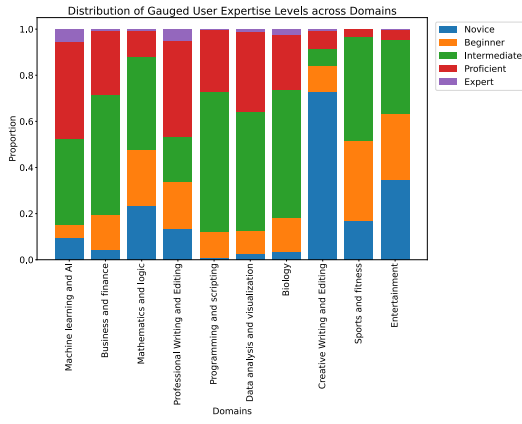


Figure 5: Barplot showing the distribution of Gauged User Expertise on different domains of Copilot conversations.

#### A.4 Reproducibility

Owing to privacy concerns, the Copilot conversations used in our study cannot be made public. However, to ensure the generalisability of our expertise classifier, we rerun our expertise classifier on a randomly sampled set of 6400 conversations from the WildChat dataset (Zhao et al., 2024). We observe a similar distribution of user, gauged user and LLM expertise labels on this sampled set of conversations. However, it is important to note that, conversations in the WildChat dataset reflect a technological bias, since the service was hosted on Hugging Face Spaces, predominantly attracting developers or individuals connected to the IT domain. As such, it may not fully represent the diversity of conversations that naturally occur with chat-based models. In contrast, Copilot conversations span 25 topically diverse domains and includes a more representative sample of conversations from the general population.

#### A.5 Prompts

We release the prompts used for expertise classification. All the prompts were human validated before being used for classification. For human validation of the measured expertise labels, three of the authors independently labeled both the user and LLM portions of 50 Bing Copilot conversations each (total of 300 assessments: the user and the LLM side of 150 conversations) as "Beginner", "Intermediate", or "Expert". We then compared these labels to a corresponding 3-class version of the final classifier, noting 70% or greater agreement in all cases. The 3-class classifier was used for human validation, as discriminating among the five classes in the full classifier was very challenging for the human labelers.

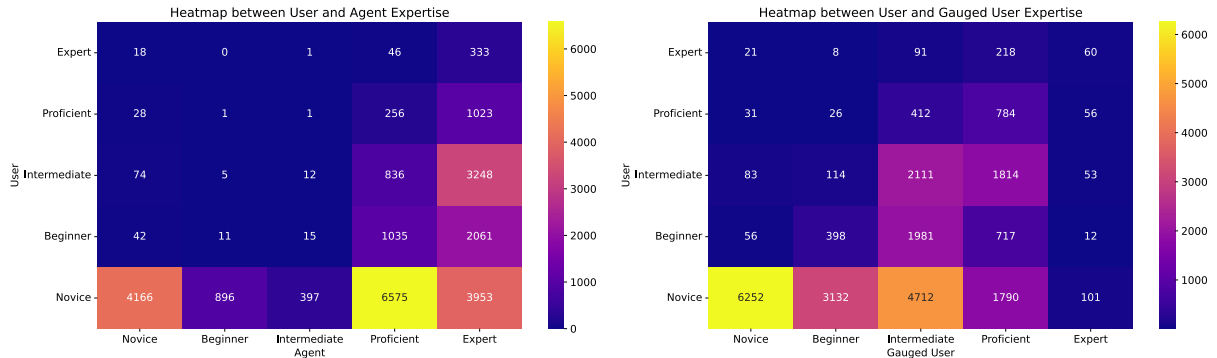


Figure 6: Heatmaps between User and AI expertise (left) and User and Gauged User expertise (right).

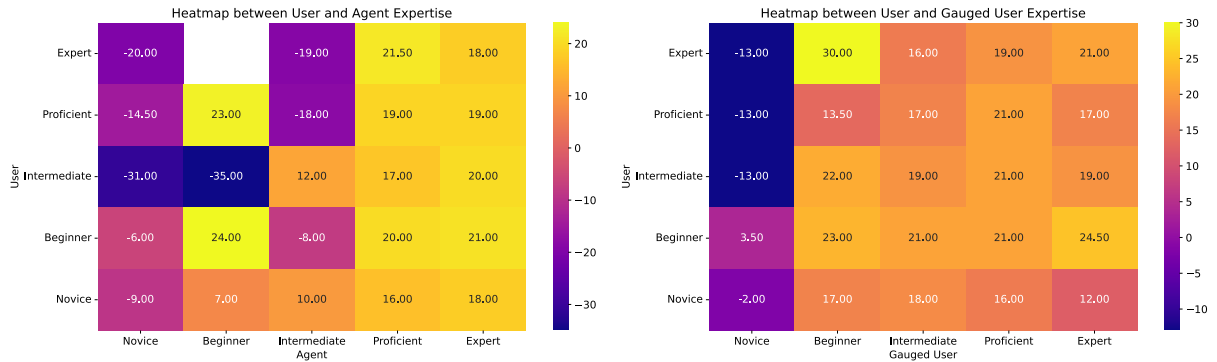


Figure 7: Heatmaps between User and AI expertise (left) and User and Gauged User expertise (right) with density as median SAT scores.

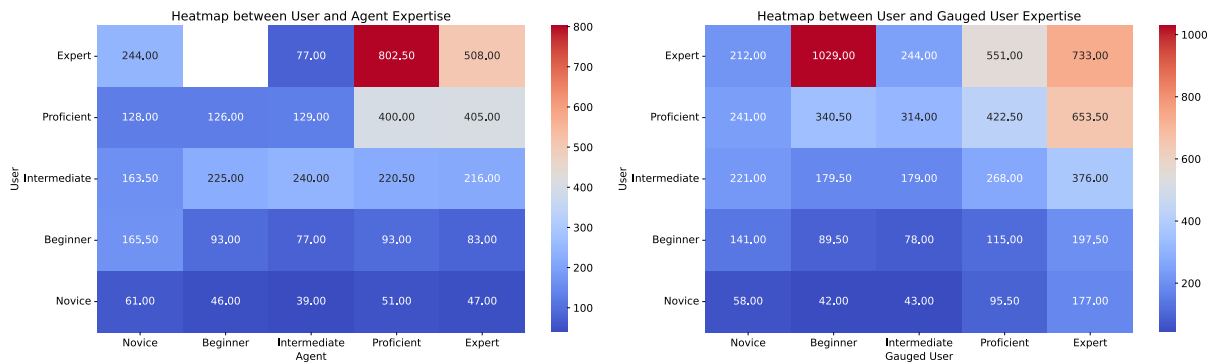


Figure 8: Heatmaps between User and AI expertise (left) and User and Gauged User expertise (right) with density as median number of words in user turns from a conversation.



Figure 9: Prompt for classifying the User Expertise based on the user only turns of Copilot conversations.



Figure 10: Prompt for classifying the AI Expertise based on the AI only turns of Copilot conversations.



Figure 11: Prompt for predicting the Gauged User Expertise based on the AI only turns of Copilot conversations.