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

Shramay Palta, Nirupama Chandrasekaran,
Rachel Rudinger and Scott Counts

Microsoft

Motivation

- LLMs have become an integral part of our day-to-day lives.
- Users interact with LLMs for a variety of complex tasks across several topical domains.
- Are all users the same?
- Can all users digest the same complexity of information?

Research Question

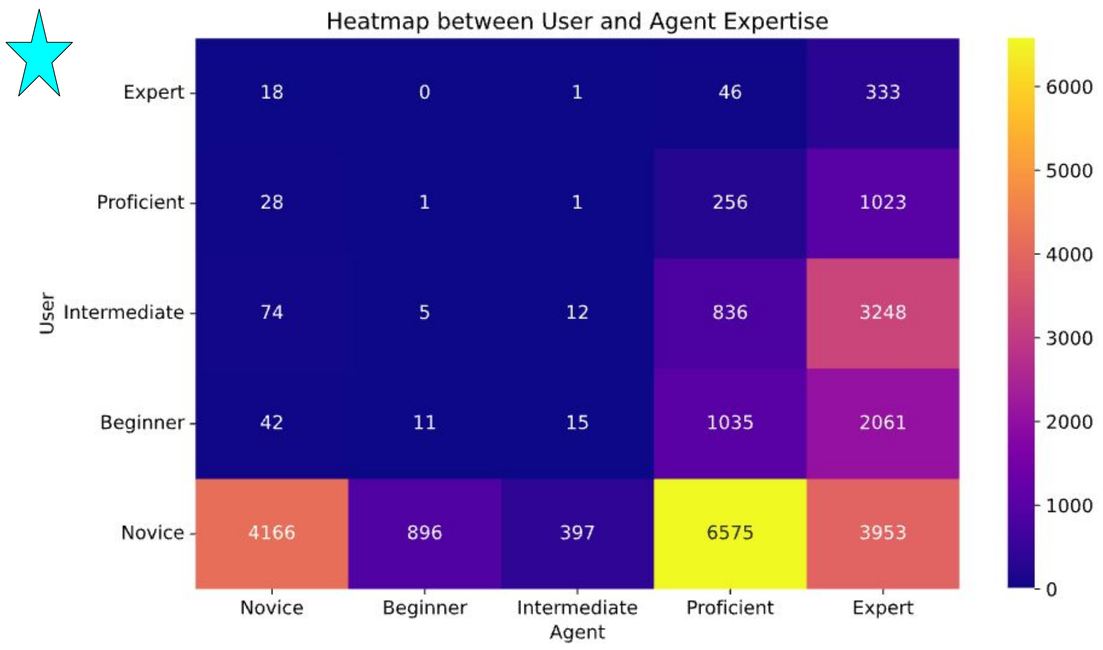
- What is the ideal expertise level of the LLM, and what are the consequences of any misalignment between the user and the LLM on domain expertise?

What we do

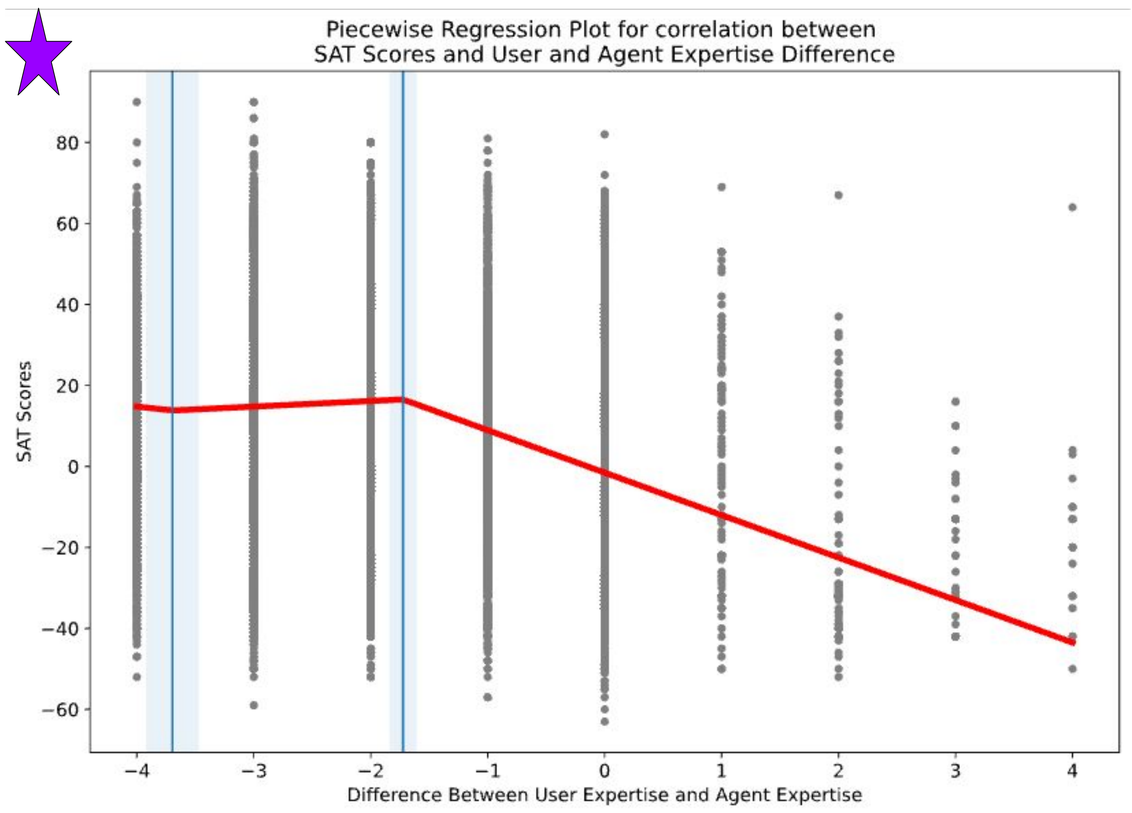
- Analyse 25000+ conversations from Bing Copilot Chat from June 2024.
- Introduce a prompt-based classifier to determine three types of expertise labels on a 5 point ordinal scale:
  - User Expertise
  - Agent Expertise
  - Gauged User Expertise
- Measure impacts of expertise (mis)alignment on user experience using metrics such as User Satisfaction (SAT) Score and Task Complexity.

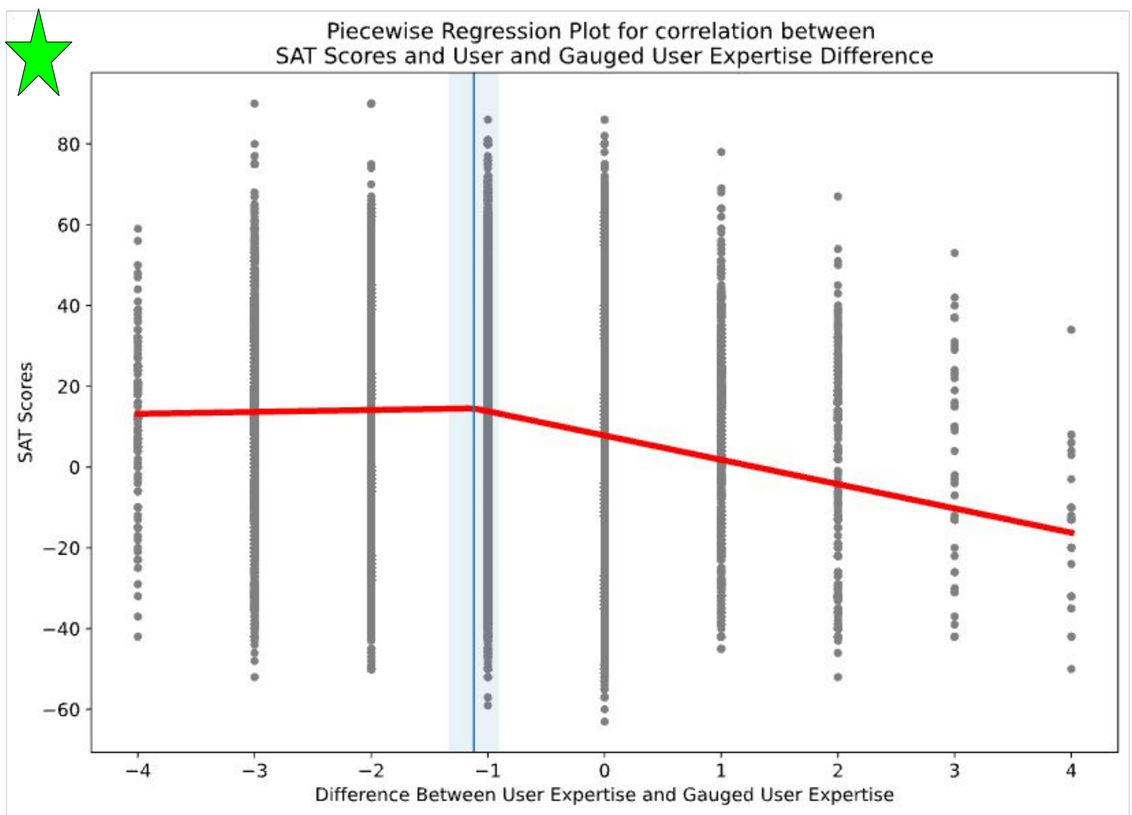
Key Takeaways

- AI is not “Proficient” or “Expert” in more than 20% of the conversations.
- AI tends to underestimate or overestimate the user expertise in most conversations.
- Low AI expertise negatively impacts the SAT Score.
- Underestimating the user expertise hurts the SAT Score.



More than 1 in 5 cases, the Agent is not “Proficient” or “Expert”!





When the gap between User Expertise and Agent Expertise (top) or the Gauged User Expertise (bottom) is large, the user satisfaction (SAT) score drops significantly

References

[1] Interpretable User Satisfaction Estimation for Conversational Systems with Large Language Models (Lin et al., ACL 2024)
[2] The Use of Generative Search Engines for Knowledge Work and Complex Tasks(Suri et al., 2024)