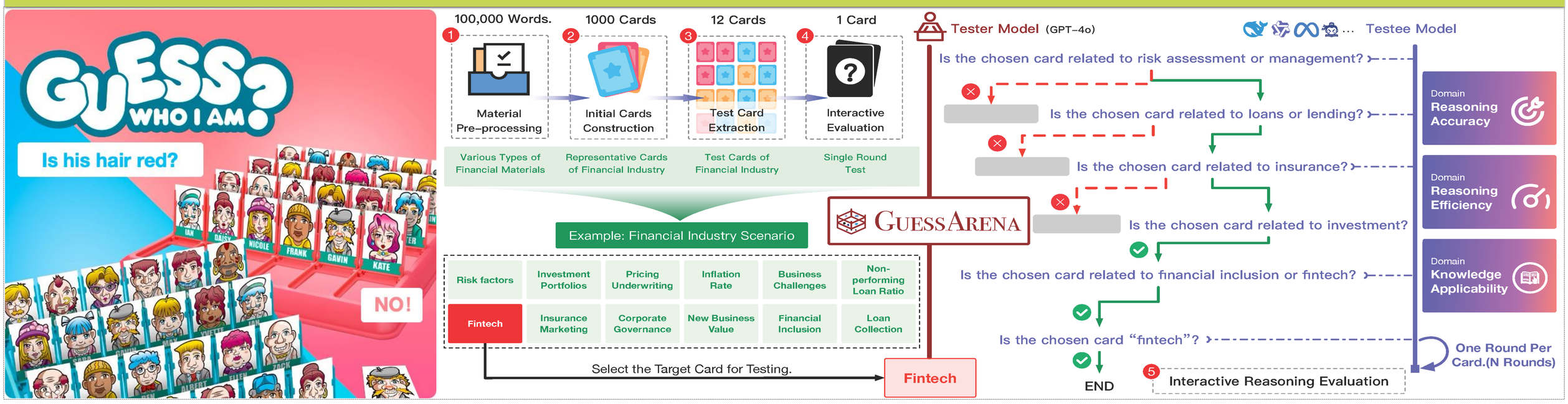
GuessArena: Guess Who I Am? A Self-Adaptive Framework for Evaluating LLMs in Domain-Specific Knowledge and Reasoning

Qingchen Yu[1]* Zifan Zheng[2]* Ding Chen[3]* Simin Niu[4] Bo Tang[1] Feiyu Xiong[1] Zhiyu Li[1]†

[1]MemTensor (Shanghai) Technology Co., Ltd. [2]University of Sydney

[3]Research Institute of China Telecom [4]Renmin University of China

Framework of GuessArena



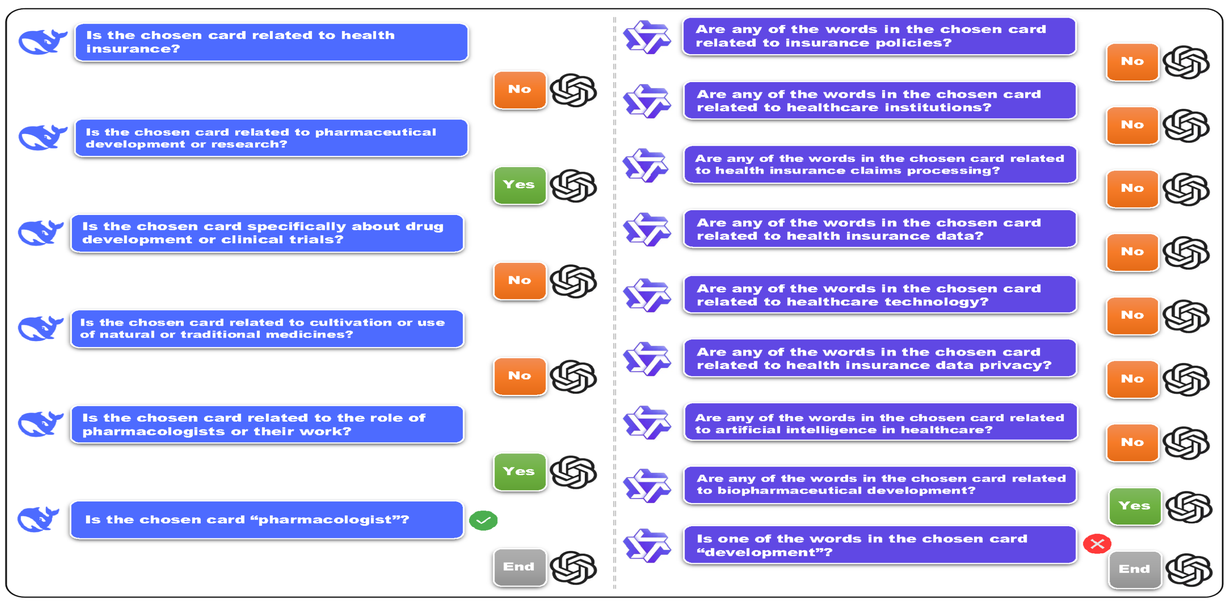
Introduction

Problem: Evaluating large language models (LLMs) in specialized domains (e.g., finance, medicine, law) is challenging. Existing static benchmarks often fail to measure an LLM's true reasoning and up-to-date knowledge capabilities comprehensively.

Our Solution: We introduce **GuessArena**, a novel self-adaptive framework for evaluating LLMs.

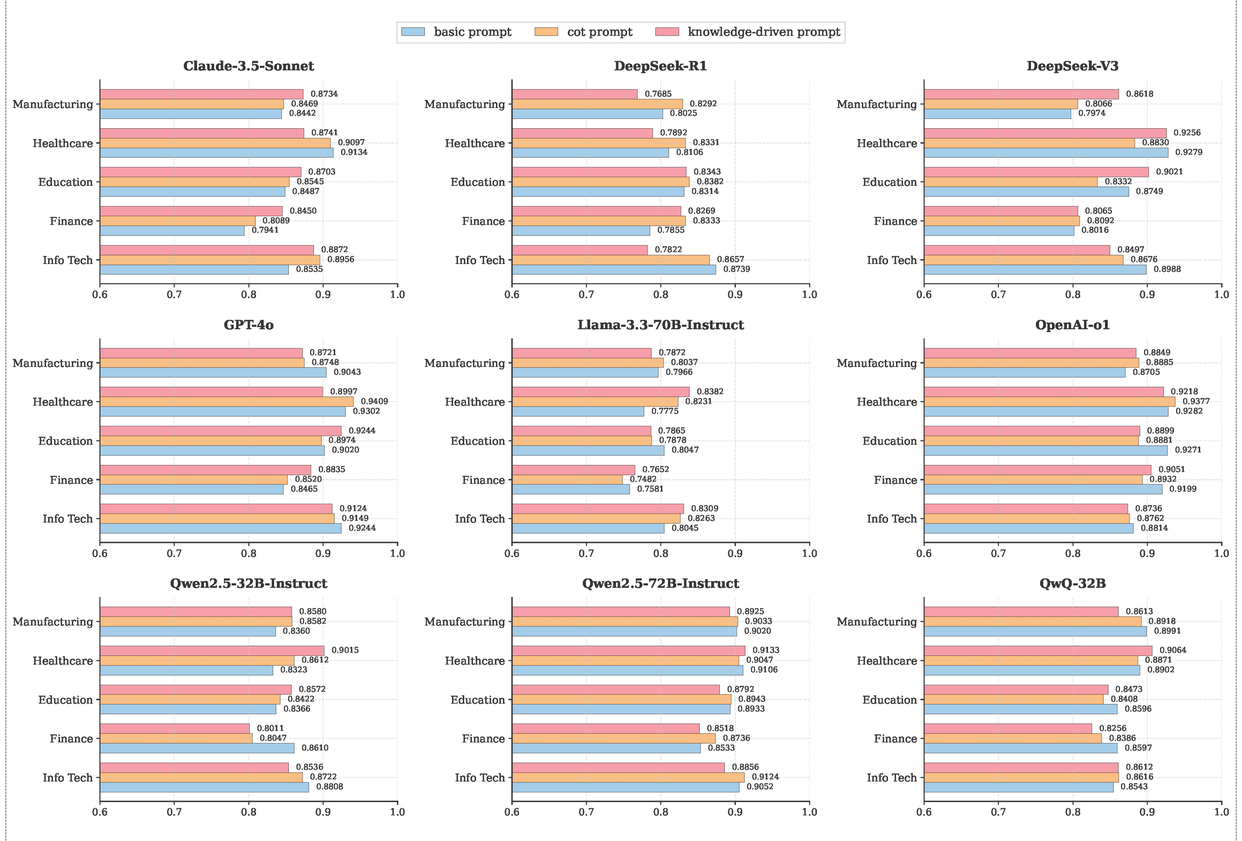
Core Idea: GuessArena gamifies the evaluation process through a "Guess Who I Am?" task. An evaluator LLM must identify a target card by probing it with dynamically generated questions, forcing the target to reveal its domain-specific knowledge and reasoning abilities.

Contribution: GuessArena provides a more dynamic, robust, and fine-grained method to assess and differentiate the capabilities of various LLMs in high-stakes domains.



Interactive Guessing Trajectory: A Healthcare Scenario Example

Results



Cross-domain GuessArena scores (higher is better) for nine LLMs under three prompting strategies.

Conclusion

A Novel Adaptive Framework: We introduce GuessArena, an evaluation framework inspired by the "Guess Who I Am?" game. It is designed to overcome the key limitations of traditional static benchmarks, such as their lack of adaptability for diverse domains and vulnerability to data contamination.

Automated and Scalable Pipeline: The framework uniquely combines two core components: 1) A dynamic knowledge modeling stage that automatically extracts evaluation "cards" from unstructured documents, significantly reducing the cost and effort of creating domain-specific tests. 2) An interactive reasoning assessment that uses a multi-turn game to quantitatively measure an LLM's questioning strategies and logical effectiveness.

Effective Differentiation of LLM Abilities: Our experiments across five vertical domains (finance, healthcare, etc.) demonstrated that GUESSARENA effectively distinguishes the performance of nine state-of-the-art LLMs. Crucially, the framework can pinpoint whether a model's weakness stems from its reasoning ability (performance improves with CoT prompts) or its domain knowledge (performance improves with knowledge-driven prompts).

Scan the QR code to access our code on GitHub.