TurtleBench: Evaluating Top Language Models via Real-World Yes/No Puzzles

Qingchen Yu[1] Shichao Song[2] Ke Fang[1] Yunfeng Shi[3] Zifan Zheng[1] Hanyu Wang[2]
Simin Niu[2] Zhiyu Li[1] Yanfang Chen[2] Xiaoqing Zheng[2]
[1]Institute for Advanced Algorithms Research (Shanghai)
[2]Renmin University of China [3]5Y Capital

Background & Motivation

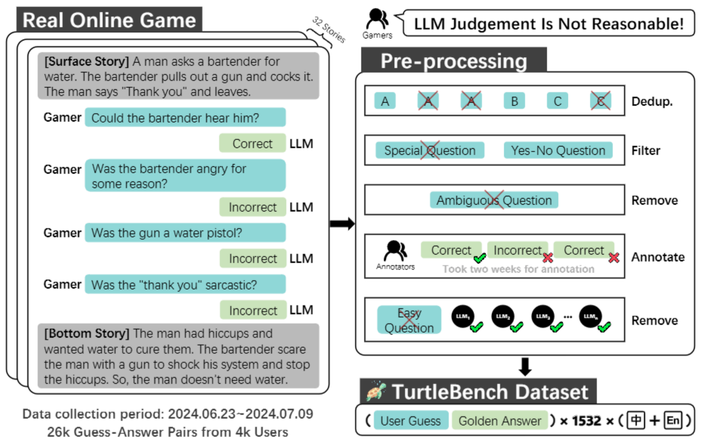
- **Limitations of Static Datasets:** Existing LLM benchmarks rely on static data, which fails to capture dynamic human-AI interactions and risks data contamination (cheating).
- **Knowledge Dependency:** Traditional benchmarks heavily rely on background knowledge, making it difficult to isolate and measure pure logical reasoning capabilities.
- **Costly & Biased Evaluations:** Dynamic evaluations using strong LLMs (e.g., GPT-4-as-a-judge) or human efforts introduce bias and are too costly for large-scale application.

Our Solution: TurtleBench

We propose **TurtleBench**, a dynamic evaluation benchmark based on real-world "Turtle Soup" (Yes/No) puzzles.

- **Zero Background Knowledge Required:** All necessary information is self-contained within the "Surface Story" and "Bottom Story." It strictly assesses pure logical reasoning.
- **Objective & Quantifiable:** Evaluates reasoning through explicit, objective "Correct / Incorrect" ground truths, eliminating subjective scoring bias.

Dataset Construction



We developed an online Turtle Soup puzzle platform to collect genuine human-AI interaction data:

- **Data Collection:** Gathered 26,000 guess-answer pairs from over 4,000 real users interacting with 32 distinct stories.
- **Pre-processing:** Deduplicated queries, filtered out non-Yes/No questions, and removed ambiguous guesses.
- Annotation: Conducted rigorous multi-turn manual annotation over two weeks to classify labels into binary "Correct" or "Incorrect" (treating "Unknown" as Incorrect for stability).
- **Final Dataset:** 1,532 high-quality, accurately annotated bilingual (Chinese & English) evaluation pairs. *(Insert Figure 1: TurtleBench Construction here)

Main Results

Table 2: Zero-Shot Evaluation Results[5]

Model	Story-Level Avg. Acc.	Overall Acc. ↑	F1 Score
GPT-4o	88.05%	87.66%	0.8501
Claude-3.5-Sonnet	87.63%	87.53%	0.8436
OpenAI o1-preview	84.65%	84.40%	0.8071
Qwen-2-72B	83.62%	82.90%	0.7741
Moonshot-v1-8k	82.80%	82.05%	0.7619
Llama-3.1-405B	82.39%	81.79%	0.8114
Deepseek-V2.5	80.48%	79.77%	0.7368
Llama-3.1-70B	79.44%	78.33%	0.7340
OpenAI o1-mini	73.66%	73.69%	0.6480

Table 3: Two-Shot Evaluation Results

Model	Story-Level Avg. Acc.	Overall Acc. ↑	F1 Score
Claude-3.5-Sonnet	90.00%	89.49%	0.8729
GPT-4o	87.89%	87.92%	0.8521
Qwen-2-72B	85.85%	85.12%	0.8152
Moonshot-v1-8k	84.71%	84.07%	0.8039
Llama-3.1-405B	82.20%	81.72%	0.8061
Deepseek-V2.5	81.70%	80.68%	0.7723
Llama-3.1-70B	79.52%	79.37%	0.7713

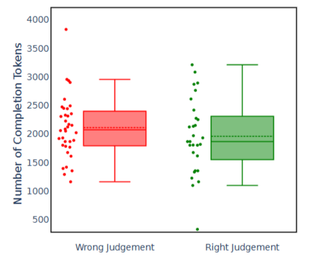


Figure 4: Completion Token Lengths for Wrong and Right Judgments of o1-preview

Conclusion

TurtleBench offers a reliable, dynamic, and objective framework to assess LLMs' real-world reasoning and understanding capabilities, mitigating data contamination risks.