Is Your Language Model Ready for Monetization Decisions?

Anonymous ACL submission

Abstract

Language language models (LLMs) are increasingly used in monetization systems for user understanding, advertisement selection, and recommendation evaluation. However, existing benchmarks primarily focus on general knowledge and open-domain reasoning, offering limited coverage of the domain-specific challenges present in real-world monetization settings. We introduce MonBench, a high-quality multi-task benchmark designed to evaluate language models in realistic monetization scenarios. MonBench is formulated as multiple-choice questions spanning user intent understanding and relationship reasoning, and is constructed from billions of production-scale monetization data points with expert annotations and controlled difficulty to ensure reliability and discriminative power. We evaluate approximately 20 state-of-the-art LLMs on MonBench and uncover substantial performance gaps relative to general-purpose benchmarks, exposing persistent limitations in monetization-specific reasoning and intent understanding. We further analyze sensitivity to evaluation and prompting choices, including answer order, output formats, instruction fine-tuning, reasoning, and in-context learning, and conduct per-task error analyses to identify reasoning-intensive monetization challenges that remain difficult even for the strongest models. Together, these findings establish MonBench as a rigorous benchmark and offer practical guidance for developing and evaluating LLMs in real-world monetization systems.

1 Introduction

Language language models (LLMs) play a central role in modern online platforms by enabling large-scale monetization through advertising, recommendation, and sponsored content (Jia et al., 2025; Xu et al., 2024; Lu et al., 2025; Tang et al., 2025a). Monetization-driven decision making underlies many user-facing services, including search engines and e-commerce platforms, where systems must simultaneously balance user experience, advertiser objectives, and platform revenue. In recent years, LLMs have been widely applied to monetization-related tasks such as ad relevance modeling (Liu et al., 2025c; Tang et al., 2025b), user intent understanding (Zhou et al., 2025; Chang et al., 2025; Wang et al., 2024a), creative and product matching (Mehrdad et al., 2024; Chen et al., 2025b; Cheng et al., 2024), and recommendation (Wang et al., 2024b; Liu et al., 2025b; Zhao et al., 2024). These systems operate in highly dynamic environments with heterogeneous entities, stakeholders, and objectives, making monetization one of the most complex and business-critical application domains for machine learning.

Recent research has explored the use of LLMs in e-commerce related domains, with benchmarks and datasets primarily centered on online shopping and user–product interactions (Zhang et al., 2025; Reddy et al., 2022a; Jin et al., 2023; Chen et al., 2025a; Jin et al., 2024). These efforts have advanced tasks such as product understanding, query interpretation, review analysis, and shopping assistance, and have demonstrated the potential of LLMs to handle multi-task and few-shot learning settings. However, existing benchmarks largely abstract away the commercial decision context inherent to monetization systems. In particular, they do not model advertisement evaluation or the trade-offs between user relevance and business objectives, and platform constraints. Moreover, many prior datasets rely on synthetic constructions or proxy tasks that simplify real-world decision processes, limiting their ability to faithfully evaluate model behavior in production-like scenarios. As a result, current benchmarks provide limited insight into how LLMs perform in real monetization settings, leaving a critical gap between academic evaluation and practical deployment.

Monetization differs fundamentally from

general-purpose and shopping-focused benchmarks due to its unique commercial decision context. Monetization tasks require accurate understanding of user intent while simultaneously reasoning under relevance, economic objectives, and business or policy constraints—capabilities that are largely absent from existing general benchmarks. Prior benchmarks predominantly focus on shopping scenarios and user–product interactions, where the primary goal is matching users with products. In contrast, monetization systems must evaluate and recommend advertisements, assess creative appropriateness, and balance user satisfaction with advertiser and platform objectives, which demand qualitatively different reasoning and evaluation setups. As a result, existing shopping benchmarks fail to capture the core challenges faced by real-world monetization systems.

To address the limitations of existing benchmarks, we introduce MonBench, a high-quality multi-task benchmark for evaluating LLMs in real-world monetization decision contexts (Figure 1). MonBench is constructed from billions of production data points sourced from multiple leading search engines, with explicit difficulty control to ensure the benchmark remains challenging as LLM capabilities continue to improve. All questions are annotated by domain experts. The benchmark covers a diverse set of monetization tasks, including ad relevance assessment, user intent alignment, and product understanding. All tasks are reformulated into a unified multiple-choice format, enabling systematic comparison across model architectures and training paradigms. Using MonBench, we conduct extensive evaluations on a wide range of LLMs, comparing model performance across monetization tasks. We also perform detailed per-task analyses, including identifying reasoning-intensive tasks and examining the impact of few-shot prompting strategies. Our results highlight key strengths and limitations of current LLMs in monetization settings, providing insights into their reasoning behavior, generalization capability, and suitability for deployment in practical monetization systems.

To summarize, our main contributions are:

- We introduce a high-quality multi-task evaluation benchmark for real-world monetization scenarios, formulated as multiple-choice questions derived from billions of production data points, with difficulty control and expert annotations to ensure reliability.
- We evaluate approximately 20 state-of-the-art LLMs on the benchmark, analyzing model performance as well as sensitivity to answer order shuffling, output formats, instruction fine-tuning, reasoning capabilities, and in-context learning.
- We conduct detailed per-task analyses, including error analysis, to identify reasoning-intensive tasks and to study the effects of few-shot prompting, providing insights into the unique challenges of monetization-oriented evaluation.

2 Related Work

2.1 E-commerce Datasets and Benchmarks

Previously, e-commerce datasets and benchmarks have primarily focused on user–product interactions and shopping assistance. Earlier datasets target individual tasks (Yang et al., 2021; Reddy et al., 2022b), including session-based recommendation and text generation (Jin et al., 2023), e-commerce concept understanding (Chen et al., 2025a), or factual knowledge grounding via knowledge graphs (Liu et al., 2025a). More recent benchmarks expand task scope through richer supervision or multi-task formulations. Shopping-MMLU (Jin et al., 2024) unifies diverse shopping tasks into a large-scale benchmark, demonstrating the potential of LLMs as general shop assistants, while ShoppingComp (Tou et al., 2025) evaluates LLM-powered shopping agents on realistic product retrieval, expert-level report generation, and safety-critical decision making.

Despite these advances, existing benchmarks remain fundamentally shopping-centric, evaluating relevance, assistance quality, or safety in isolation. In contrast, monetization systems operate in a distinct decision context, requiring multi-objective and constraint-aware reasoning that jointly balances user intent, advertiser objectives, platform policies, and revenue optimization. These economically grounded and policy-constrained decision processes are largely absent from current e-commerce benchmarks. Motivated by this gap, we introduce a monetization-focused multi-task benchmark that enables systematic evaluation of LLMs in real-world monetization settings.

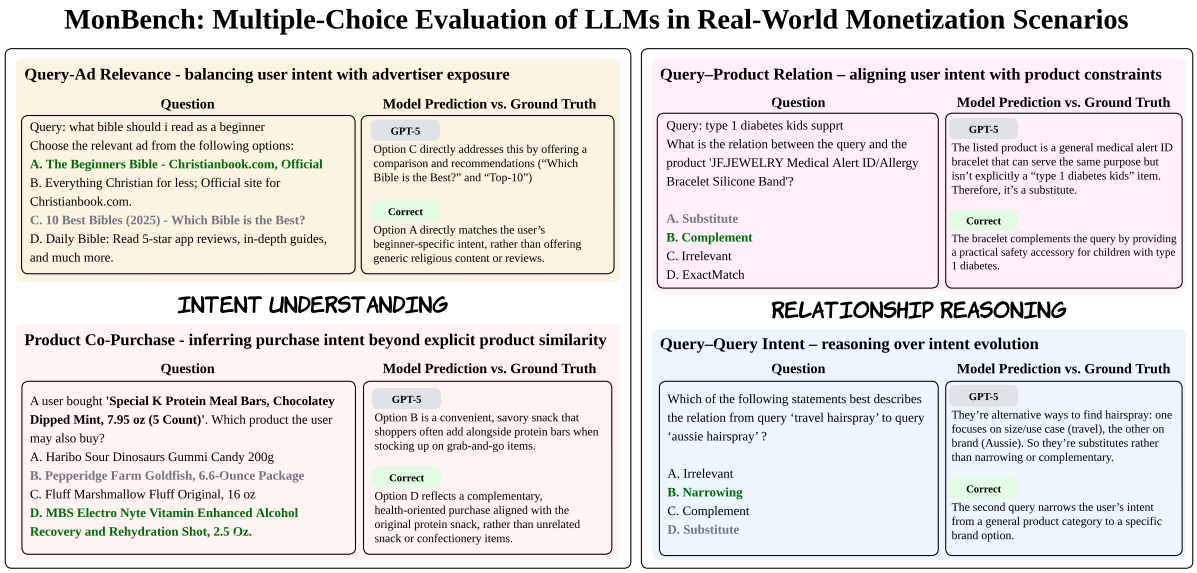


Figure 1: MonBench evaluates language models on monetization decision-making by covering both explicit intent understanding and relational reasoning among queries, advertisements, and products, capturing real-world challenges unique in monetization scenarios.

2.2 Web & Shopping Agent Benchmarks

Recent work has explored agent-based benchmarks to evaluate LLMs in interactive shopping and web environments. Benchmarks such as Mind2Web (Deng et al., 2023), WebArena (Zhou et al., 2024), WebMall (Peeters et al., 2025), WebShop (Yao et al., 2023), and Shopping-Bench (Wang et al., 2025a) evaluate LLM agents on end-to-end task completion over deterministic or simulated web interfaces, emphasizing navigation, tool use, and multi-step interaction. Other works, including DeepShop (Lyu et al., 2025) and Amazon-Bench (Zhang et al., 2025), further assess LLMs' ability to perform realistic shopping tasks on real-world e-commerce platforms, while OPeRA (Wang et al., 2025b) focuses on simulating human shopping behaviors through structured observation–action trajectories.

While these benchmarks advance research on web navigation and shopping agents, they remain largely shopping-centric and focus on interaction fidelity or task completion. They do not explicitly capture the multi-objective, constraint-aware decision making required in monetization systems, where models must jointly reason about user intent, advertiser goals, platform policies, and revenue trade-offs. Our benchmark is complementary to these agent evaluations, as robust monetization understanding and reasoning are prerequisites for reliable performance in complex interactive decision-making in real-world settings.

3 MonBench Overview

This section presents the design of MonBench, which consists of four tasks organized into two categories and constructed from billions of real-world data.

3.1 Data Source

As shown in Table 1, MonBench is constructed from a combination of real-world commercial data sources spanning advertising and e-commerce scenarios from multiple leading search engines. Together, these sources enable MonBench to reflect practical monetization contexts involving queries, advertisements, and products across multiple platforms. All data are anonymized and do not contain identifiable user information or internal proprietary data.

3.2 Data Collection Pipeline

MonBench is constructed through a structured data collection pipeline tailored to the characteristics of each task category, as illustrated in Figure 2. The pipeline integrates automated filtering, stratified sampling and difficulty control, and expert validation to ensure data quality, balanced difficulty, and coverage across diverse monetization scenarios.

Query–Ad Relevance. For query–ad relevance, we stratify queries by intent and retrieve candidate advertisements. To control task difficulty, we

Task Group	Task	Total Size	Few-shot Size	Validation Size	Test Size	Original Data Scale	Data Source
Intent Understanding	Query–Ad Relevance	442	5	40	397	Billions	Search Engine 1
	Product Co-Purchase	375	5	34	336	Thousands	ShoppingMMLU
Relationship Reasoning	Query–Query Intent	593	5	53	535	Thousands	ShoppingMMLU
	Query–Product Relationship	1482	5	134	1343	Billions	Search Engine 1 & 2, ShoppingMMLU

Table 1: Overview of task groups, dataset sizes, original data scale, and data sources in MonBench.

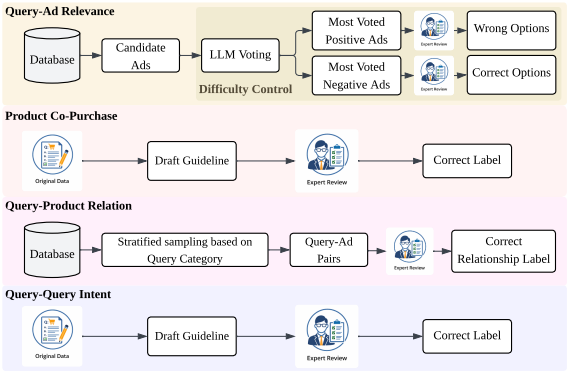


Figure 2: Overview of the data collection and annotation pipeline for MonBench.

apply an LLM-based voting scheme to identify high-confidence relevance predictions, which are selectively inverted by human annotators to surface challenging cases. The verified examples are then reformulated into multiple-choice questions, with relevant ads as correct options and irrelevant ads as wrong options.

Query–Product Relationship. For this task, we first perform stratified sampling over the database based on query categories to ensure balanced coverage of different intent types. Sampled query–product pairs are then formed and reviewed by human experts, who assign the correct semantic relationship labels (e.g., exact match, substitute, complement, or irrelevant).

Query–Query Intent and Product Co-Purchase. For the query–query intent and product co-purchase tasks, we leverage original multiple-choice data sourced from ShoppingMMLU (Jin et al., 2024). We first draft task-specific annotation guidelines to standardize labeling criteria. Human experts then review each example under these guidelines and assign the final correct labels, ensuring consistency and high annotation quality across tasks.

3.3 Data Quality Control

All data in MonBench is annotated by a group of expert human labelers following carefully designed and extensively discussed annotation guidelines. These guidelines specify task objectives, label definitions, and representative examples, and were iteratively refined through pilot annotation rounds and group discussions. During annotation, labelers applied the shared guidelines consistently across tasks, with ambiguous cases reviewed and resolved through majority voting to reduce annotation noise and bias.

3.4 Few-shot Setup and Data Split

We adopt a 5-shot prompting setting, where up to five labeled examples are provided in-context to the model. For evaluation, the data is split into validation and test sets with a 1:10 ratio (Table 1). The validation set is used for prompt selection and hyperparameter tuning (e.g., instruction phrasing and formatting choices), while the held-out test set is used exclusively for final performance reporting to ensure an unbiased evaluation.

4 Experiments

4.1 Experimental Setup

Metrics. We evaluate model performance using accuracy, as all tasks in MonBench are formulated as multiple-choice questions with a single correct answer. Prior work has shown that LLMs' performances on multiple-choice evaluations can be sensitive to the ordering of answer options as well as the required output format (e.g., predicting the option label, the option content, or both) (Zheng et al., 2024; Pezeshkpour and Hruschka, 2023; Gupta et al., 2024). To mitigate these sources

of variance, we evaluate each question under all 24 permutations of the four answer choices and three output formats, resulting in 72 inference runs per question. Final results are reported as the mean accuracy across these 72 runs, along with the corresponding 95% confidence interval, providing a more stable and robust estimate of model performance.

Evaluated models. We evaluate a diverse set of LLMs spanning different architectures, scales, and reasoning capabilities. Specifically, we include models from the GPT family (OpenAI et al., 2024), the DeepSeek series (DeepSeek-AI et al., 2025a,b), and the Qwen3 model series (Yang et al., 2025). For Qwen3, we evaluate multiple model sizes and configurations, including base models as well as variants with explicit reasoning enabled (think) and disabled (nothink). This selection allows us to systematically analyze the effects of model scale and reasoning mechanisms on performance across all tasks in MonBench.

4.2 Overall Performance

Reasoning-enabled models consistently dominate the MonBench leaderboard, as reported in Table 2. GPT-5.2 and GPT-5-Reasoning achieve the highest overall accuracy and exhibit the strongest, most consistent performance across all tasks, highlighting the effectiveness of explicit reasoning capabilities in monetization-oriented decision making.

We observe a clear scaling trend across models: performance generally improves with increasing model size, and reasoning-enabled variants tend to outperform non-reasoning models at comparable parameter scales. However, scale alone does not fully determine performance. Notably, DeepSeek-v3 (671B) achieves accuracy comparable to Qwen3-32B-nothink, despite a substantial difference in model size, suggesting that architectural and training choices can partially compensate for scale in these tasks.

Based on overall accuracy, models can be grouped into three performance tiers.

- Tier 1 (>80% accuracy) includes GPT-5.2 and gpt-5-reasoning, followed by gpt-5-chat. Models in this tier consistently achieve the highest accuracy across both relevance selection and relationship classification tasks, indicating strong general capability and robust cross-task reasoning.
- Tier 2 (72–80% accuracy) comprises GPT-5.1, DeepSeek-R1, GPT-4o, GPT-5-mini, DeepSeek-V3, and larger Qwen3 variants (e.g., Qwen3-32B and Qwen3-14B think). Performance within this tier is largely ordered by model size, with reasoning-enabled models generally outperforming their non-reasoning counterparts.
- Tier 3 (<72% accuracy) consists primarily of smaller Qwen3 models ($\leq$ 8B) and mid-sized non-reasoning variants (e.g., Qwen3-14B non-think). While reasoning variants continue to provide measurable gains, they do not fully offset the performance degradation caused by reduced model scale.

Overall, these results highlight the complementary roles of model scale and reasoning capability in monetization-focused benchmarks: reasoning improves performance consistently, but sufficient scale remains critical for achieving top-tier accuracy.

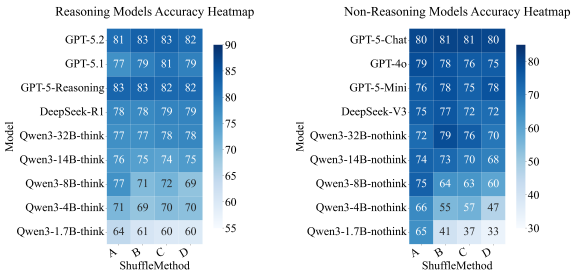
4.3 Sensitivity Analysis

Shuffling Order Analysis. LLMs are known to be sensitive to answer option ordering in multiple-choice evaluations (Pezeshkpour and Hruschka, 2023; Zheng et al., 2024; Gupta et al., 2024). To analyze this effect, we evaluate model performance across all 24 permutations of four answer choices and report the average accuracy when the correct option appears at positions A–D in Figure 3. We find that accuracy varies systematically with option placement, revealing non-negligible positional bias across model families. This bias is most pronounced in smaller models, which strongly favor earlier options (e.g., A or B), leading to large performance fluctuations. Chat-oriented models exhibit moderate sensitivity, while reasoning-enabled models are substantially more robust, with relatively stable accuracy across permutations. These results underscore the importance of controlling for option-order effects and motivate the use of permutation-averaged metrics for fair model comparison.

Return Format Analysis. We further analyze model sensitivity to return format in Figure 4, examining how different output styles affect multiple-choice accuracy. We observe a clear dependence on model size: smaller models exhibit substantially

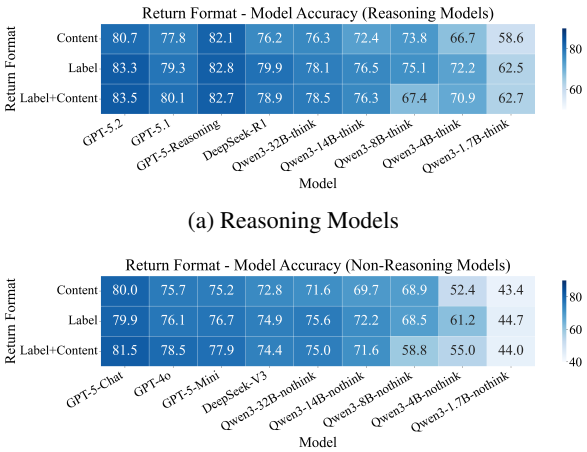
Model Type	Model	Intent Understanding		Relationship Classification		Avg. (95% CI)
		Query-Ad Relevance	Product Co-Purchase	Query-Product Relation	Query-Query Intent	
Reasoning	GPT-5.2	**86.88**	88.08	69.38	**85.74**	82.52 ± 0.36
Reasoning	GPT-5-Reasoning	85.92	**88.13**	**70.50**	85.45	82.50 ± 0.15
Non-Reasoning	GPT-5-Chat	85.83	86.86	65.05	84.17	80.48 ± 0.35
Reasoning	GPT-5.1	84.69	86.54	61.71	83.34	79.07 ± 0.69
Reasoning	DeepSeek-R1	83.37	84.64	61.71	83.67	78.35 ± 0.30
Reasoning	Qwen3-32B-think	81.85	82.93	61.46	84.28	77.63 ± 0.33
Non-Reasoning	GPT-4o	83.98	87.00	54.16	82.04	76.80 ± 0.69
Non-Reasoning	GPT-5-Mini	85.66	86.49	56.15	77.99	76.57 ± 0.70
Reasoning	Qwen3-14B-think	83.78	84.49	48.75	83.24	75.07 ± 0.40
Non-Reasoning	Qwen3-32B-nothink	77.12	79.15	56.79	83.27	74.08 ± 1.50
Non-Reasoning	DeepSeek-V3	81.66	82.87	51.93	79.65	74.03 ± 0.94
Reasoning	Qwen3-8B-think	82.04	80.79	57.92	67.54	72.07 ± 1.29
Non-Reasoning	Qwen3-14B-nothink	78.98	81.17	43.78	80.61	71.13 ± 1.15
Reasoning	Qwen3-4B-think	80.52	78.48	46.40	74.32	69.93 ± 0.43
Non-Reasoning	Qwen3-8B-nothink	77.24	74.45	46.61	63.34	65.41 ± 2.42
Reasoning	Qwen3-1.7B-think	75.25	68.10	41.99	59.67	61.25 ± 0.88
Non-Reasoning	Qwen3-4B-nothink	73.72	68.90	35.50	46.65	56.19 ± 2.98
Non-Reasoning	Qwen3-1.7B-nothink	57.95	43.85	30.17	44.18	44.04 ± 5.31

Table 2: Performance of reasoning and non-reasoning language models on MonBench tasks. Results are reported as accuracy (%) with 95% confidence intervals.



(a) Reasoning Models (b) Non-Reasoning Models

Figure 3: Accuracy heatmaps for reasoning and non-reasoning models under different shuffle methods (correct answer at A–D).



(a) Reasoning Models

(b) Non-Reasoning Models

Figure 4: Accuracy heatmaps for reasoning and non-reasoning models across different return formats.

higher variance across return formats, while larger models remain comparatively stable with minimal performance fluctuations. In addition, reasoning-enabled models are the most robust to changes in return format, showing consistently low variance across output styles. In contrast, non-reasoning models display greater performance variability, indicating a higher reliance on the specific format of model outputs.

4.4 Insights

4.4.1 General Knowledge Transfer

To assess how well general knowledge transfers to monetization-specific tasks, we compare model performance on MonBench with results on MMLU-Redux (Gema et al., 2025), as shown in Figure 5. Overall, MonBench is consistently more challenging, with lower absolute accuracy across models, reflecting its higher difficulty and stronger domain-specific requirements. While performance rankings remain largely consistent, where models that perform well on MMLU-Redux also tend to perform better on MonBench, we observe a pronounced divergence across model scales. In particular, smaller models exhibit relatively competitive performance on the general benchmark but experience substantial degradation on MonBench. This gap suggests that monetization tasks impose unique challenges that are not adequately captured by general knowledge benchmarks, disproportionately affecting models with limited capacity.

4.4.2 Effects of Instruction Fine-tuning

In Figure 6, we compare the performance of Qwen3 base models and instruction fine-tuned non-thinking variants on MonBench. We find that instruction fine-tuned models consistently outperform their base counterparts across all model scales, with particularly large gains on monetization-

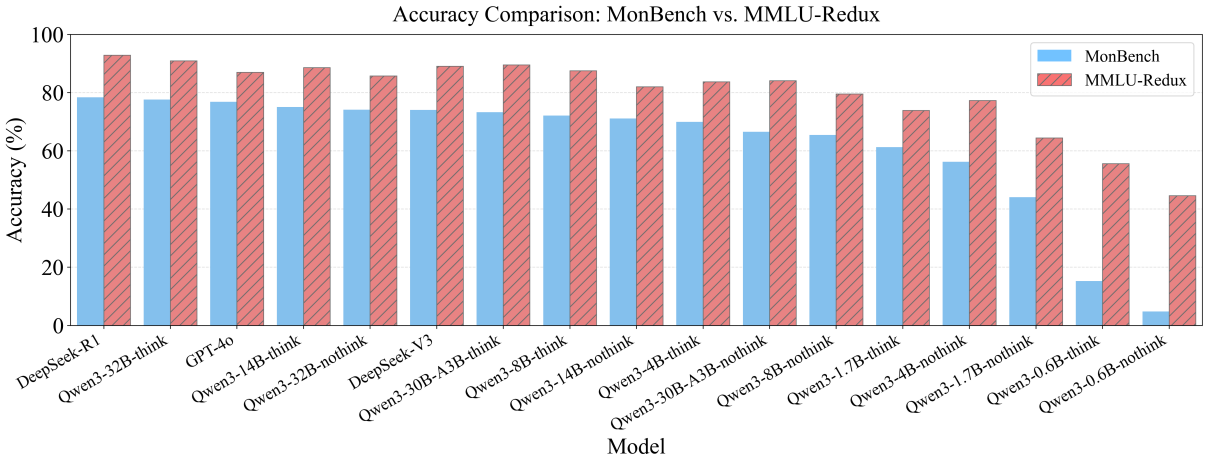
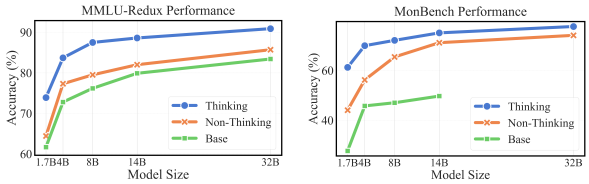


Figure 5: Performance on MonBench and MMLU-Redux shows that monetization tasks have unique challenges.

focused multi-task evaluations. These results suggest that instruction fine-tuning is critical for aligning LLMs with complex, application-driven objectives, and that effective instruction tuning is essential for strong real-world decision-making performance beyond model scale or reasoning capability alone.



(a) MMLU-Redux (b) MonBench

Figure 6: Performance comparison on MMLU-Redux and MonBench for Qwen3 base models, thinking and non-thinking models. Qwen3-32B base is not released and is therefore missing from this plot.

4.4.3 Effects of Reasoning

In Figure 6, we compare the performance of Qwen3 non-thinking models and thinking variants on MonBench. We observe that reasoning-enabled models consistently outperform both base and non-reasoning variants across all model scales. The benefit of reasoning is particularly pronounced for smaller models (e.g., Qwen3 1.7B and 4B), where reasoning yields substantial performance improvements on monetization tasks. In contrast, larger models (e.g., Qwen3 8B, 14B, and 32B) exhibit more modest gains, suggesting that increased model capacity partially compensates for the absence of explicit reasoning. These results indicate that reasoning mechanisms are especially valuable for improving performance in lower-capacity models on complex, application-driven tasks.

4.4.4 Effects of In-context Learning

We analyze the impact of in-context learning across different task types and observe heterogeneous effects.

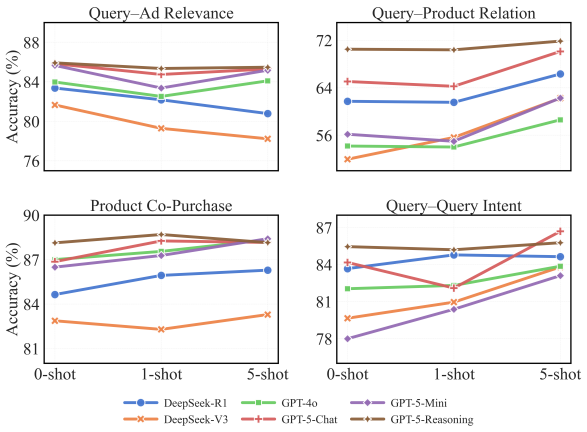


Figure 7: Results of in-context learning (0-, 1-, and 5-shot) on MonBench for LLMs.

For GPT models and DeepSeek models (over 100B), tasks involving explicit user intent, such as query–ad relevance and product co-purchase, exhibit smaller or mixed performance improvements with additional in-context examples. In contrast, reasoning-intensive tasks, including query–product relationship and query–query intent classification, benefit more substantially from in-context learning. These results, as shown in Figure 7, suggest that in-context examples are particularly effective for tasks requiring relational inference and semantic reasoning, while offering limited gains for tasks dominated by direct relevance matching.

For Qwen3 reasoning models, few-shot prompt-

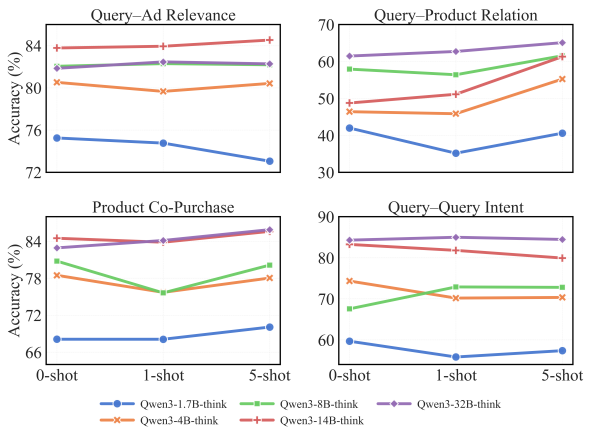


Figure 8: Results of in-context learning (0-, 1-, and 5-shot) on MonBench for Qwen3 reasoning models.

ing yields mixed or marginal performance improvements, with accuracy largely saturating at 0–1 shot across tasks (Figure 8). This pattern suggests that explicit reasoning capabilities already capture most of the benefits typically provided by in-context examples, leaving limited room for additional gains from further prompting. As a result, additional in-context examples offer minimal incremental improvement for reasoning-enabled models.

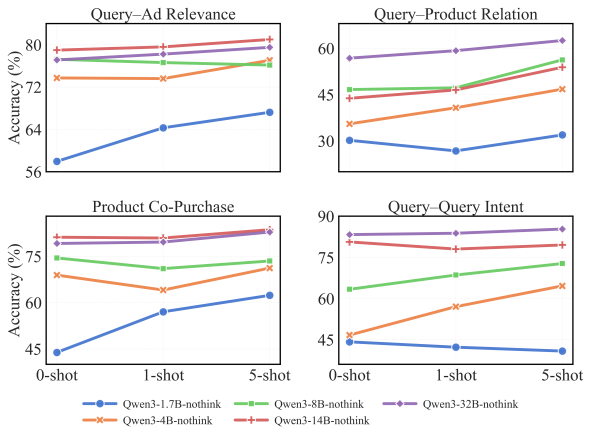


Figure 9: Results of in-context learning (0-, 1-, and 5-shot) on MonBench for Qwen3 non-reasoning models.

For Qwen3 non-reasoning models, few-shot prompting leads to mixed or modest performance gains across tasks (Figure 9). While additional in-context examples can improve accuracy, these gains are insufficient to close the gap with larger or reasoning-enabled models, particularly on reasoning-intensive tasks. This suggests that in-context learning alone cannot fully compensate for the absence of explicit reasoning mechanisms in non-reasoning models.

4.5 Error Analysis

We conduct an error analysis on three GPT models using GPT-5.2 as an automated judge to categorize model failures and verified by humans, in Figure 10. For each task and model, we randomly sample 500 erroneous predictions and label them into four categories: 1) Reasoning errors, where the model understands the intent and has access to relevant information but fails in logical inference or decision-making; 2) Lack of specific knowledge, where errors stem from missing or incorrect domain knowledge; 3) user intent misinterpretation, where the model misunderstands the user's goal or task requirements; and 4) other rrrors, which include formatting issues, instruction non-compliance, or ambiguous cases. Case examples and more details are provided in Appendix D.2. Across all tasks, reasoning errors emerge as the dominant failure mode, particularly for reasoning-intensive tasks such as Query–Product Relationship and Query–Query Intent. While stronger reasoning-enabled models reduce knowledge-related and intent misinterpretation errors, these results indicate that improving reasoning robustness remains the primary opportunity for further performance gains.

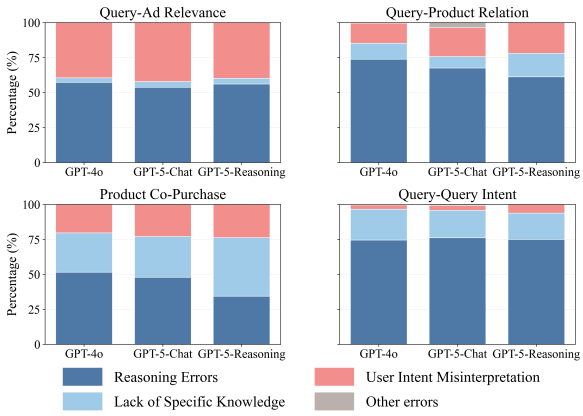


Figure 10: Error category breakdown for three GPT models across four tasks.

5 Conclusion

In this paper, we present MonBench, a real-world benchmark for evaluating language models on monetization-related tasks involving user intent understanding and relationship classification. Our results highlight the benefits of reasoning-enabled models and the complementary role of model scale. We hope MonBench supports future research on implicit intent, multi-step reasoning, and agent-based decision making in monetization applications.

6 Limitations

While MonBench covers realistic monetization scenarios, it has two main limitations: it focuses on explicitly expressed user intent rather than implicit signals from user behavior or context, and it evaluates single-step decisions without requiring multi-step reasoning or planning. Future work could extend the benchmark to implicit-intent settings and agent-centric tasks involving sequential decision making.

References

Zongyu Chang, Feihong Lu, Ziqin Zhu, Qian Li, Cheng Ji, Zhuo Chen, Hao Peng, Yang Liu, Ruifeng Xu, Yangqiu Song, Shangguang Wang, and Jianxin Li. 2025. Bridging the gap between llms and human intentions: Progresses and challenges in instruction understanding, intention reasoning, and reliable generation. *Preprint*, arXiv:2502.09101.

Haibin Chen, Kangtao Lv, Chengwei Hu, Yanshi Li, Yujin Yuan, Yancheng He, Xingyao Zhang, Langming Liu, Shilei Liu, Wenbo Su, and Bo Zheng. 2025a. Chineseecomqa: A scalable e-commerce concept evaluation benchmark for large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5311–5321, New York, NY, USA. Association for Computing Machinery.

Junyi Chen, Lu Chi, Siliang Xu, Shiwei Ran, Bingyue Peng, and Zehuan Yuan. 2025b. Hllm-creator: Hierarchical llm-based personalized creative generation. *Preprint*, arXiv:2508.18118.

Zhu Cheng, Wen Zhang, Chih-Chi Chou, You-Yi Jau, Archita Pathak, Peng Gao, and Umit Batur. 2024. E-commerce product categorization with LLM-based dual-expert classification paradigm. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 294–304, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Preprint*, arXiv:2306.06070.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. 2024. Changing answer order can decrease mmlu accuracy. *Preprint*, arXiv:2406.19470.

Weitao Jia, Shuo Yin, Zhoufutu Wen, Han Wang, Zehui Dai, Kun Zhang, Zhenyu Li, Tao Zeng, and Xiaohui Lv. 2025. Summa: A multimodal large language model for advertisement summarization. *Preprint*, arXiv:2508.20582.

Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Xiao Cheng, Rahul Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang Tang, Bing Yin, and Xianfeng Tang. 2023. Amazon-m2: A multilingual multi-locale shopping session dataset for recommendation and text generation. *Preprint*, arXiv:2307.09688.

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, Haodong Wang, Zhengyang Wang, Wenju Xu, Jingfeng Yang, Qingyu Yin, Xian Li, Priyanka Nigam, Yi Xu, Kai Chen, and 3 others. 2024. Shopping mmlu: A massive multi-task online shopping benchmark for large language models. *Preprint*, arXiv:2410.20745.

Langming Liu, Haibin Chen, Yuhao Wang, Yujin Yuan, Shilei Liu, Wenbo Su, Xiangyu Zhao, and Bo Zheng. 2025a. Eckgbench: Benchmarking large language models in e-commerce leveraging knowledge graph. *Preprint*, arXiv:2503.15990.

Qidong Liu, Xiangyu Zhao, Yuhao Wang, Yejing Wang, Zijian Zhang, Yuqi Sun, Xiang Li, Maolin Wang, Pengyue Jia, Chong Chen, Wei Huang, and Feng Tian. 2025b. Large language model enhanced recommender systems: A survey. *Preprint*, arXiv:2412.13432.

Tongtong Liu, Zhaohui Wang, Meiyue Qin, Zenghui Lu, Xudong Chen, Yuekui Yang, and Peng Shu. 2025c. Real-time ad retrieval via llm-generative commercial intention for sponsored search advertising. *Preprint*, arXiv:2504.01304.

Qingyi Lu, Haotian Lyu, Jiayun Zheng, Yang Wang, Li Zhang, and Chengrui Zhou. 2025. Research on e-commerce long-tail product recommendation mecha-

nism based on large-scale language models. *Preprint*, arXiv:2506.06336.

Yougang Lyu, Xiaoyu Zhang, Lingyong Yan, Maarten de Rijke, Zhaochun Ren, and Xiuying Chen. 2025. Deepshop: A benchmark for deep research shopping agents. *Preprint*, arXiv:2506.02839.

Navid Mehrdad, Hrushikesh Mohapatra, Mossaab Bagdouri, Prijith Chandran, Alessandro Magnani, Xunfan Cai, Ajit Puthenputhussery, Sachin Yadav, Tony Lee, ChengXiang Zhai, and Ciya Liao. 2024. Large language models for relevance judgment in product search. *Preprint*, arXiv:2406.00247.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ralph Peeters, Aaron Steiner, Luca Schwarz, Julian Yuya Caspary, and Christian Bizer. 2025. Webmall – a multi-shop benchmark for evaluating web agents [technical report]. *Preprint*, arXiv:2508.13024.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022a. Shopping queries dataset: A large-scale esci benchmark for improving product search. *Preprint*, arXiv:2206.06588.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022b. Shopping queries dataset: A large-scale esci benchmark for improving product search. *Preprint*, arXiv:2206.06588.

Brian Jay Tang, Kaiwen Sun, Noah T. Curran, Florian Schaub, and Kang G. Shin. 2025a. Ads that talk back: Implications and perceptions of injecting personalized advertising into llm chatbots. *Preprint*, arXiv:2409.15436.

Tian Tang, Zhixing Tian, Zhenyu Zhu, Chenyang Wang, Haiqing Hu, Guoyu Tang, Lin Liu, and Sulong Xu. 2025b. Lref: A novel llm-based relevance framework for e-commerce search. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 468–475. ACM.

Huaixiao Tou, Ying Zeng, Cong Ma, Muzhi Li, Minghao Li, Weijie Yuan, He Zhang, and Kai Jia. 2025. Shoppingcomp: Are llms really ready for your shopping cart? *Preprint*, arXiv:2511.22978.

Jiangyuan Wang, Kejun Xiao, Qi Sun, Huaipeng Zhao, Tao Luo, Jian Dong Zhang, and Xiaoyi Zeng. 2025a. Shoppingbench: A real-world intent-grounded shopping benchmark for llm-based agents. *Preprint*, arXiv:2508.04266.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024a. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.

Qi Wang, Jindong Li, Shiqi Wang, Qianli Xing, Runliang Niu, He Kong, Rui Li, Guodong Long, Yi Chang, and Chengqi Zhang. 2024b. Towards next-generation llm-based recommender systems: A survey and beyond. *Preprint*, arXiv:2410.19744.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton, and Dakuo Wang. 2025b. Opera: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. *Preprint*, arXiv:2506.05606.

Wei Xu, Jue Xiao, and Jianlong Chen. 2024. Leveraging large language models to enhance personalized recommendations in e-commerce. *Preprint*, arXiv:2410.12829.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2021. Mave: A product dataset for multi-source attribute value extraction. *Preprint*, arXiv:2112.08663.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2023. Webshop: Towards scalable real-world web interaction with grounded language agents. *Preprint*, arXiv:2207.01206.

Xianren Zhang, Shreyas Prasad, Di Wang, Qiuhai Zeng, Suhang Wang, Wenbo Yan, and Mat Hans. 2025. A functionality-grounded benchmark for evaluating web agents in e-commerce domains. *Preprint*, arXiv:2508.15832.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Qianrui Zhou, Hua Xu, Yifan Wang, Xinzhi Dong, and Hanlei Zhang. 2025. Llm-guided semantic relational reasoning for multimodal intent recognition. *Preprint*, arXiv:2509.01337.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

A Task Examples

In this section, we present an example question for each task, along with the system prompt and model's answer for references.

A.1 Query-Ad Relevance

Given a user query and contextual signals (e.g., location), the model selects the most relevant advertisement from multiple candidates. This task is challenging because relevance must balance user intent satisfaction with advertiser exposure, rather than optimizing for either objective alone.

A.2 Query-Product Relation

Given a user query and a candidate product, the model classifies their relationship (e.g., exact match, substitute, complement, or irrelevant). This task tests the model's ability to align user intent with product-specific constraints rather than relying on surface semantic overlap.

A.3 Product Co-Purchase

Given a purchased product, the model identifies another product the user may buy together. Correct predictions require inferring implicit purchase intent and real-world co-purchase behavior beyond explicit product similarity.

A.4 Query-Query Intent

Given two user queries from the same session, the model determines their semantic relationship (e.g., narrowing, reformulation, substitute, or irrelevant). Success requires reasoning over intent evolution across successive queries instead of treating queries as independent inputs.

B Dataset Details

B.1 System Prompts for Different Return Formats

To study the impact of output constraints on model behavior and evaluation stability, we evaluate models under multiple return formats, each defined by a corresponding system prompt. These prompts control how the model is instructed to produce its final answer, ranging from minimal label-only outputs to label outputs accompanied by natural-language explanations.

Specifically, we consider two return formats. The *Label-only* format requires the model to select a single answer choice and output only the corresponding label in a fixed, machine-readable format. The *Label + Explanation* format additionally allows the model to provide a brief natural-language explanation justifying its selection, enabling analysis of reasoning behavior while preserving structured outputs.

Table 7 summarizes the system prompts used for each return format. All prompts are designed to be concise, task-agnostic, and consistent across tasks to ensure fair comparison across models and evaluation settings.

C Few-shot Prompting Examples

In addition to zero-shot evaluation, we assess model performance under few-shot prompting to examine whether limited in-context supervision improves task understanding and decision-making. Few-shot examples are provided directly in the system prompt as demonstrations of the expected input–output format, while remaining task-agnostic and free of test data leakage.

We consider two few-shot settings: *1-shot* and *5-shot*. In the 1-shot setting, the prompt includes a single illustrative example that demonstrates the task structure and the required output format. In the 5-shot setting, five diverse examples are included to provide broader coverage of task variations. All examples are sampled from held-out data that does not overlap with the validation or test sets.

Table 8 presents representative 1-shot and 5-shot examples used in our experiments. These examples are identical across models to ensure fair comparison.

Field	Content
Question	**User Query:** electric chargers warrensburg mo **Location:** Kansas City, MO **Market:** en-us **Ad Options:** **A.** Local EV Charger Installers – Enter Zip & Get Estimates. Free EV charging station cost estimates from local installers. **B.** Just Enter Your Zip Code – EV Charging Stations – Electricians. Find top-rated EV charging station experts in your area. **C.** EV Home Charging Station Installers. Compare quotes and save. **D.** EV Charging Station Install – Get Up To Four Quotes Free. Customized quotes from trusted contractors.
System Prompt	You will be given a user query along with the location and market where the query was made. Using this information, choose the most relevant advertisement from the provided options. Output the answer in the format `<Label>A/B/C/D</Label>`.
Model Answer	**Prediction:** `<Label>A</Label>` **Explanation:** The query indicates intent to find EV charger installation services in Warrensburg, MO. Option A explicitly targets local EV charger installers, prompts ZIP-code-based matching, and offers free estimates, making it the most relevant choice despite the broader Kansas City location context.
Correct Answer	`B`

Table 3: Example of a Query–Ad Relevance multiple-choice question in MonBench, illustrating the need to balance user intent and location with advertiser offerings.

D Experiment Details

D.1 List of Language Model Studied

Model Overview. We summarize the evaluated models below, along with a brief description of their design characteristics and performance tendencies on MonBench.

Evaluated Models. We evaluate a diverse set of proprietary and open-weight large language models spanning different architectures, scales, and reasoning configurations.

- **GPT-5.2**: A proprietary large language model released by OpenAI, representing the latest generation of the GPT-5 family with enhanced reasoning capabilities.
- **GPT-5-Reasoning**: A reasoning-focused variant of GPT-5 explicitly optimized for multi-step inference and structured decision-making.
- **GPT-5-Chat**: A conversational variant of GPT-5 designed for general-purpose dialogue and instruction following without explicit reasoning mechanisms.
- **GPT-5.1**: An earlier generation GPT-5-series model with built-in reasoning capabilities, serving as a strong pre-GPT-5.2 baseline.
- **GPT-5-Mini**: A compact proprietary model from OpenAI designed for efficiency and lower-latency deployment while retaining strong instruction-following ability.
- **GPT-4o**: A proprietary multimodal language model from OpenAI that supports text and vision inputs, primarily optimized for general-purpose understanding.
- **DeepSeek-R1**: A reasoning-oriented open-weight language model released by DeepSeek, designed to support explicit chain-of-thought style inference.
- **DeepSeek-V3**: A non-reasoning open-weight language model from DeepSeek, intended for general-purpose language understanding and generation.
- **Qwen3-32B-think**: A 32B-parameter reasoning-enabled model from the Qwen3 family.
- **Qwen3-32B-nothink**: A 32B-parameter non-reasoning variant of Qwen3, sharing the same parameters as its reasoning counterpart but without explicit chain of thought thinking enabled.
- **Qwen3-14B-think**: A 14B-parameter reasoning-enabled model from the Qwen3

Field	Content
Question	**User Query:** what kills mold naturally **Product Ad:** Kilz Mold & Mildew White Flat Water-Based Alkyd Mold Killing Primer (1 qt) **Price:** $27.76 **Seller:** Walmart **Relation Options:** **A.** ExactMatch **B.** Substitute **C.** Irrelevant **D.** Complement
System Prompt	You are a helpful online shopping assistant. Please answer the following question about online shopping. Choose the option that best describes the relationship between the product and the query. Output the answer in the format `<Label>A/B/C/D</Label>`.
Model Answer	**Prediction:** `<Label>C</Label>` (Irrelevant) **Explanation:** The query seeks natural methods for killing mold, while the advertised product is a chemical-based mold-killing primer. It does not satisfy the intent for natural remedies, nor does it function as a substitute or complement to an informational query, making the product irrelevant to the user's intent.
Correct Answer	`B`

Table 4: Example of a Query–Product Relation multiple-choice question in MonBench, highlighting the challenge of aligning user intent with product constraints beyond surface topical relevance.

family, incorporating explicit chain-of-thought style reasoning during inference.

- **Qwen3-14B-nothink**: A 14B-parameter non-reasoning variant of Qwen3, sharing the same backbone architecture as its reasoning-enabled counterpart but without explicit chain-of-thought reasoning.
- **Qwen3-8B-think**: An 8B-parameter reasoning-enabled model from the Qwen3 family, designed to study reasoning behavior at smaller model scales.
- **Qwen3-8B-nothink**: An 8B-parameter non-reasoning Qwen3 model that removes explicit chain-of-thought reasoning while maintaining the same parameter scale.
- **Qwen3-4B-think**: A 4B-parameter reasoning-enabled model from the Qwen3 family, representing a compact configuration with explicit reasoning enabled.
- **Qwen3-4B-nothink**: A 4B-parameter non-reasoning variant of Qwen3, sharing the same architecture as its reasoning-enabled counterpart without explicit reasoning mechanisms.
- **Qwen3-1.7B-think**: A 1.7B-parameter reasoning-enabled Qwen3 model, included to analyze reasoning capabilities under very limited model capacity.
- **Qwen3-1.7B-nothink**: A 1.7B-parameter non-reasoning Qwen3 model, serving as a minimal-capacity baseline without explicit chain-of-thought reasoning.

D.2 Error Analysis

In this section, we present a qualitative error analysis of GPT-family models on MonBench to better understand common failure modes across monetization-oriented tasks. We categorize incorrect model predictions into four mutually exclusive error types, following a standardized labeling protocol using GPT-5.2 as an evaluation assistant. Each error case is assigned exactly one primary category based on the dominant cause of failure.

For each incorrectly answered question, the evaluation assistant is provided with the input question, the correct answer, and the model's predicted answer, and is instructed to identify the primary error type according to the following definitions:

Reasoning Errors. Reasoning errors occur when the model has access to relevant information and correctly understands the user intent, but fails due to incorrect logical reasoning, inference, or decision-making (e.g., faulty comparisons, invalid deductions, or incorrect trade-offs). An example of such error is presented in Table 11.

Lack of Specific Knowledge. This type of error refers to errors caused by missing, incorrect, or

Field	Content
Question	**User Purchase:** Joyva Marble Halvah Bars (12 Bars) **Candidate Products:** **A.** Kirkland Signature Ito En Matcha Blend (Green Tea), 100% Japanese Green Tea Leaves (200 Tea Bags) **B.** Bassetts Liquorice Allsorts (190g) **C.** MW Polar Canned Vegetables, Cut Yams (15 oz, Pack of 24) **D.** SweetLeaf Sweetener (70-Count Packets, Pack of 3)
System Prompt	You are a helpful online shopping assistant. Please answer the following question about online shopping. Select the product that the user may also buy. Output the answer in the format `<Label>A/B/C/D</Label>`.
Model Answer	**Prediction:** `<Label>A</Label>` **Explanation:** Halvah is a rich, sweet confection commonly consumed alongside tea, making green tea a natural complementary purchase. The remaining options are either alternative candies (substitutes), unrelated pantry items, or peripheral products that do not directly complement the original purchase.
Correct Answer	`B`

Table 5: Example of a Product Co-Purchase multiple-choice question in MonBench, highlighting the challenge of inferring complementary purchase intent from real-world consumption patterns.

Field	Content
Question	**Query 1:** ilia skin tint **Query 2:** ilia foundation **Relation Options:** **A.** Substitute **B.** Complement **C.** Narrowing **D.** Irrelevant
System Prompt	You are a helpful online shopping assistant. Please answer the following question about online shopping. Choose the option that best describes the relationship between the two queries. Output the answer in the format `<Label>A/B/C/D</Label>`.
Model Answer	**Prediction:** `<Label>A</Label>` (Substitute) **Explanation:** Both queries refer to ILIA complexion base products that serve the same function. A foundation is an alternative to a skin tint (typically offering fuller coverage), making the two queries interchangeable rather than complementary or a specificity change.
Correct Answer	`B`

Table 6: Example of a Query–Query Intent multiple-choice question in MonBench, illustrating the challenge of distinguishing substitutes from other semantic relations in intent evolution.

incomplete domain-specific knowledge required to answer the question correctly. An example of such error is presented in Table 9.

User Intent Misinterpretation. This type of error arises when the model misunderstands what the user is actually asking for or what the task requires, such as confusing topical relevance with intent satisfaction or misclassifying the user's goal. An example of such error is presented in Table 10.

Other Errors. Other errors include cases that do not clearly fall into the above categories, such as instruction non-compliance, formatting issues, or ambiguous failures where no single category dominates. An example of such error is presented in Table 12.

Return Format	System Prompt
Label	You are a helpful online shopping assistant. Please answer the following question about online shopping. ** Output Format ** <Label>A/B/C/D</Label> **Valid Examples** <Label>A</Label>
Content	You are a helpful online shopping assistant. Please answer the following question about online shopping. ** Output Format ** <Answer>Narrowing/Complement/Irrelevant/Substitute</Answer> **Valid Examples** <Answer>Narrowing</Answer>
Label + Content	You are a helpful online shopping assistant. Please answer the following question about online shopping. ** Output Format ** <Label>A/B/C/D</Label> <Answer>Narrowing/Complement/Irrelevant/Substitute</Answer> **Valid Examples** <Label>A</Label> <Answer>Narrowing</Answer>

Table 7: System prompts corresponding to different return formats used in MonBench evaluation.

Few-shot Setting	Prompt
1-shot	Here are some examples to help you answer the question: A user has just made a query 'woodburn self storage' with the following location and region: ; Location: 30305, Atlanta, GA; Region: US.; The query returns the ad: ; AdTitle: Woodburn Self Storage; AdDescription: Self Storage In Woodburn – Get 100% Off Your First Month On Select Units. Which of the following statements best describes the relation between the ad and the query?; ExactMatch; Irrelevant; Complement; Substitute; <Answer>ExactMatch</Answer>
5-shot	Here are some examples to help you answer the question: Which of the following statements best describes the relation from query "waterpik" to query "waterpik pink"?; Narrowing; Irrelevant; Complement; Substitute; <Answer>Narrowing</Answer>; Which of the following statements best describes the relation from query "leather flats for women" to query "king comforter sets clearance prime"?; Complement; Irrelevant; Narrowing; Substitute; <Answer>Irrelevant</Answer>; Which of the following statements best describes the relation from query "black formal dresses for women" to query "black formal dresses for women evening party"?; Complement; Irrelevant; Narrowing; Substitute; <Answer>Narrowing</Answer>; Which of the following statements best describes the relation from query "scarf for men" to query "black couch"?; Complement; Narrowing; Irrelevant; Substitute; <Answer>Irrelevant</Answer>; Which of the following statements best describes the relation from query "bioderma" to query "cleansing oil"?; Complement; Irrelevant; Substitute; Narrowing; <Answer>Narrowing</Answer>;

Table 8: 1-shot and 5-shot prompts used for few-shot evaluation. Prompts are shown without modification to illustrate the exact context provided to models.

Field	Content
Error Type	*Lack of Specific Knowledge*
Question	**User Query:** glass block window privacy film **Location:** St. Louis, MO **Market:** en-us **Options:** **A.** AdTitle: Frosted Privacy Film for Glass Windows; AdDescription: Our technology helps you to make the right choice and save money. Quick and easy!. We compared all the small details for you and checked hundreds of customers' reviews.; AdURL: bestchoice.com **B.** AdTitle: Buy Window Stickers; AdDescription: Browse Our Wide Selection Of Stained Glass, Frosted, Etch, & Patterned Films. Order Today!; AdURL: www.wallpops.com **C.** AdTitle: Get Window Film Applied - Free Quotes - Find a Pro Near You; AdDescription: Free Quotes from Residential Window Tinting Services Near You.. Free Quotes from Window Film Installation Contractors Near You.; AdURL: www.estimates123.com/Window-Tinting **D.** AdTitle: Film On Glass For Privacy - Control Your Privacy; AdDescription: One-stop shop for all your smart glass needs. Discover more about our innovative solutions. Smart Glass Products Company. Best Prices in the US. Get Your Free Estimate in 24 hours!; AdURL: www.smartglasstech.us/smart/film
Correct Answer	**D**
Model Answer	**B**
Model Explanation	The user's query suggests they want to purchase privacy film for glass block windows, indicating a shopping intent rather than hiring an installer. Option B directly offers a wide selection of privacy films (frosted, etched, patterned) suitable for windows. Option A is a generic comparison site, C is for hiring installers, and D focuses on smart glass, which is less relevant to simple privacy film needs

Table 9: An error case categorized as *Lack of Specific Knowledge* on the Query-Ad Relevance task. The model selects option B based on general topical match, while the correct answer is D.

Field	Content
Error Type	*User Intent Misinterpretation*
Question	**User Query:** wholesale air conditioners for sale **Ad Offer:** Goodman GSX140361 3 Ton 14 SEER AC Condenser **Price:** $1608.12 **Seller:** Skip The Warehouse **Options:** (A) Complement; (B) ExactMatch; (C) Irrelevant; (D) Substitute
Correct Answer	**D** (*Substitute*)
Model Answer	**B** (*ExactMatch*, `<Label>B</Label>`)
Model Explanation	The model interprets the query as a general intent to purchase an air conditioner and treats a single retail product listing as an exact match. However, the query explicitly indicates a wholesale intent, implying bulk purchasing or wholesale-level offerings, making the individual product listing a substitute rather than an exact match.

Table 10: An error case categorized as *User Intent Misinterpretation* on the Query–Product Relation task. The model fails to distinguish wholesale intent from individual retail purchase intent.

Field	Content
Error Type	*Reasoning Error*
Question	**Query Pair:** "toddler toothbrush age 2–4" → "u shaped toothbrush kids" **Options:** (A) Irrelevant; (B) Complement; (C) Substitute; (D) Narrowing
Correct Answer	**D** (*Narrowing*)
Model Answer	**C** (*Substitute*, `<Label>C</Label>`)
Model Explanation	The model treats the two queries as alternative ways to find a children's toothbrush, focusing on their shared target demographic. However, the second query specifies a particular toothbrush design within the broader age-based category, representing a narrowing of intent rather than a substitute relationship.

Table 11: An error case categorized as a *Reasoning Error* on the Query–Query Intent task. The model fails to recognize intent refinement and incorrectly labels a narrowing relationship as substitution.

Field	Content
Error Type	*Other Errors (Label–Answer Mismatch)*
Question	**User Query:** how to get funds for going back to college at 36 **Location:** Kealakekua, Hawaii, US **Ad:** FAFSA – Online Colleges for Adults (top10.com) **Options:** (A) Complement; (B) Substitute; (C) ExactMatch; (D) Irrelevant
Correct Answer	**D** (*Irrelevant*)
Model Answer	**A** (*Complement*, `<Label>A</Label>`)
Model Explanation	The model correctly explains that the advertisement does not address the user's intent to obtain funding for returning to college and therefore concludes that the ad is irrelevant. However, it outputs an incorrect label that does not align with its explanation, indicating an instruction-following or label-mapping error rather than a semantic misunderstanding.

Table 12: An error case categorized as *Other Errors*, where the model's explanation aligns with the correct answer but the predicted label is incorrect due to a label–answer mismatch.