CoPA: Benchmarking Personalized Question Answering with Data-Informed Cognitive Factors

Hang Su[1,2], Zequn Liu[2,*], Chen Hu[2], Xuesong Lu[1,*], Yingce Xia[2,*], Zhen Liu[2]

[1] East China Normal University, China. [2] Zhongguancun Academy, China.
s-sh25@bza.edu.cn, liuzequn@bza.edu.cn, huchen@bza.edu.cn,
xslu@dase.ecnu.edu.cn, xiayingce@bza.edu.cn, pt-lz@bza.edu.cn

Abstract

While LLMs have demonstrated remarkable potential in Question Answering (QA), evaluating personalization remains a critical bottleneck. Existing paradigms predominantly rely on lexical-level similarity or manual heuristics, often lacking sufficient data-driven validation. We address this by mining *Community-Individual Preference Divergence* (CIPD), where individual choices override consensus, to distill six key personalization factors as evaluative dimensions. Accordingly, we introduce CoPA, a benchmark with 1,985 user profiles for fine-grained, factor-level assessment. By quantifying the alignment between model outputs and user-specific cognitive preferences inferred from interaction patterns, CoPA provides a more comprehensive and discriminative standard for evaluating personalized QA than generic metrics. The code is available at https://github.com/bjzgcai/CoPA.

1 Introduction

Personalized Question Answering (QA) marks a paradigm shift from static, generic responses to dynamic, user-centric knowledge delivery (Nam et al., 2009; West et al., 2014; Lan et al., 2022; Xu et al., 2025). For instance, explaining "gravity" to a child requires an intuitive narrative involving Newton's apple, whereas a physicist demands a rigorous explanation grounded in mathematical formulas. Recent Large Language Models (LLMs) (Achiam et al., 2023; Team et al., 2023; Guo et al., 2025; Yang et al., 2025) have significantly advanced the field, leveraging their instruction-following and context-adaptation capabilities to dynamically generate such tailored responses. Despite the rapid advancement in LLM-driven personalized QA methods (Li et al., 2024; Liu et al., 2025a; Singh et al., 2024), establishing a reliable evaluation protocol for personalization remains a critical challenge (Salemi et al., 2024b; Kumar et al., 2024; Salemi and Zamani, 2025).

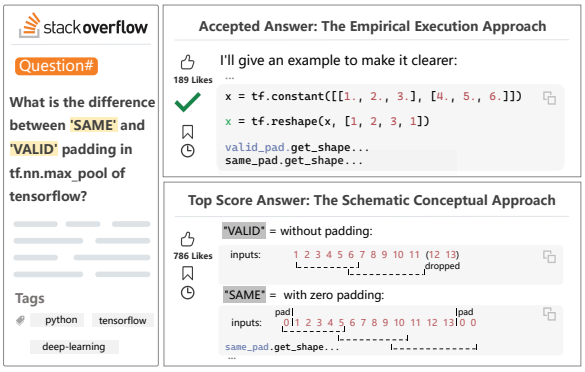


Figure 1: This Stack Overflow example illustrates Community-Individual Preference Divergence (CIPD): the user accepted an empirical, code-based solution while the community overwhelmingly favored a conceptual, schematic one.

Existing research on the evaluation of personalized generative responses generally falls into three categories. First, traditional NLP metrics (Salemi et al., 2024b; Kumar et al., 2024; Shi et al., 2025), such as BLEU (Papineni et al., 2002) and ROUGE (Papineni et al., 2002), focus on the lexical overlap between the generated text and a reference. They fail to capture actual personalized factors, as they only check for matching words rather than adaptation to the user. Second, heuristic-based evaluations (Su et al., 2025; Dai et al., 2024), attempt to compare the generated response directly to the historical content of users. Although this approach utilizes personalized information, it still relies on lexical-level similarity, failing to capture deeper user intent. Third, the "LLM-as-a-Judge" paradigm (Salemi and Zamani, 2025; Dong et al., 2024) uses LLMs to assess personalization. However, current approaches often rely on generic prompts or manually defined rules to evaluate personalization, lacking the specific criteria or de-

* Corresponding authors: Yingce Xia, Xuesong Lu, Zequn Liu.

Findings of the Association for Computational Linguistics: ACL 2026, pages 14978–15007

tailed metrics necessary for rigorous assessment. These limitations motivate us to rethink the key factors that determine personalization.

To this end, we adopt a data-driven approach to systematically mine user preferences and disentangle the essential factors of personalization from real-world online QA data. We focus on a prevalent scenario in the online QA systems: Community-Individual Preference Divergence (CIPD) (Adamic et al., 2008). Specifically, in this divergence, the answer endorsed by the community (top-voted) differs from the one adopted by the individual user (accepted), even when both are of high quality (Figure 1). Such divergence provides a natural and scalable signal for evaluating personalization.

Leveraging CIPD samples, we distilled personalized factors at scale through the following steps: First, we utilized LLMs to induce user-level decision rationales from CIPD samples. By modeling the user's historical profile alongside the contrasting answer pair, we enabled the LLM to infer the plausible cognitive rationales behind the user's specific choice. We then leveraged LLMs to summarize these rationales, eventually distilling six cognitive factors as evaluative dimensions that govern personalization assessment. Notably, although these factors were derived from StackExchange data, we demonstrate in this work that the resulting factor-based evaluation exhibit robust generalization capabilities when applied to new scenarios.

Based on these data-informed factors, we constructed a personalized QA benchmark, CoPA, to evaluate personalized question answering with LLMs. The benchmark comprises a curated set of 1,985 users from diverse domains, each associated with a historical interaction profile. For evaluation, we quantified user profiles across the six factors and adopted a 3-point Likert scale to measure the alignment between model-generated responses and these profiles.

Our contribution can be summarized as follows:
(1) We utilized LLMs to automatically conduct an in-depth exploration of the CIPD phenomenon at the user level and distilled a set of interpretable cognitive factors from large-scale online QA data.
(2) We propose CoPA, a novel factor-driven benchmark that enables quantitative evaluation of alignment between personalized content generated by LLMs and users' cognitive preferences as reflected in interaction patterns.
(3) We conducted extensive experiments on the CoPA benchmark, established comprehensive baselines, and provided in-depth quantitative and qualitative analyses, offering a solid empirical foundation for future research.

2 Analysis of CIPD Phenomenon

We investigate the CIPD phenomenon by leveraging data from StackExchange[1], a representative multi-domain Q&A platform.

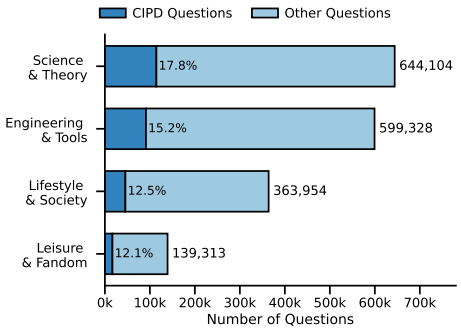
StackExchange dataset. Stack Exchange serves as a massive online Q&A network encompassing 173 distinct communities, characterized by a diverse user base of over 100 million monthly visitors. These domains can be categorized into four primary groups: **Engineering & Tools**, **Science & Theory**, **Lifestyle & Society**, and **Leisure & Fandom**, comprehensively covering the majority of real-world QA scenarios. Table 1 presents the top four domains within each category based on question volume, along with their respective distributions. On this platform, a single question elicits multiple responses, and the community votes to reflect the general consensus. Concurrently, the question asker selects a specific response using an "accepted" label to indicate his/her personal preference. This explicit distinction between community endorsement and individual choice renders Stack Exchange an ideal environment for investigating the CIPD phenomenon. A comprehensive statistical breakdown of question counts for all domains is provided in the Appendix A.

Category	Domain	Count	Pct.(%)
Engineering & Tools (Total: 1,537,431)	unix	158,149	10.3
	tex	119,815	7.8
	salesforce	85,487	5.6
	electronics	81,643	5.3
	Other	1,092,337	71.0
Science & Theory (Total: 1,652,195)	math	1,109,289	67.1
	physics	136,393	8.3
	stats	104,284	6.3
	mathematica	50,863	3.1
	Other	251,366	15.2
Lifestyle & Society (Total: 664,689)	english	96,299	14.5
	ell	79,979	12.0
	diy	44,782	6.7
	money	28,599	4.3
	Other	415,031	62.4
Leisure & Fandom (Total: 287,908)	gaming	77,043	26.8
	scifi	53,021	18.4
	travel	33,644	11.7
	rpg	31,922	11.1
	Other	92,301	32.1

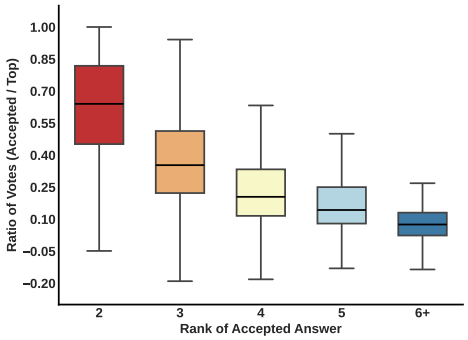
Table 1: Distribution of domains and question counts.

CIPD Questions. In this analysis, we exclude

[1] https://stackexchange.com/



(a) Proportion of CIPD questions.



(b) Rank distribution of CIPD questions.

Figure 2: Analysis of CIPD questions. (a) shows the proportion across categories, while (b) illustrates the rank distribution.

questions containing only a single answer. As illustrated in the Figure 2a, the CIPD phenomenon is pervasive across all primary categories, notably persisting even in domains typically assumed to have objective factual answers, such as Engineering & Tools. Figure 2b reveals significant disparities in vote counts between accepted and top-voted answers when divergence occurs, indicating that individual users frequently prioritize distinct criteria over community consensus. To examine whether answer timing affects CIPD, we conduct a statistical analysis of answer timestamps for CIPD questions, showing that 72.70% of accepted answers were posted no earlier than the corresponding top-voted answers. Overall, the results confirm the prevalence of the CIPD phenomenon within the StackExchange dataset, reflecting the personalized needs of individual users.

3 Factor distillation

This section outlines our method for mining personalized factors from the StackExchange dataset. Figure 3 shows the pipeline of the Factor Distillation process. First, we curated a high-quality CIPD dataset by linking user QA histories with community voting information. Second, we leveraged LLMs to extract the post-hoc decision rationales behind user choices. Finally, we aggregated these rationales to distill core factors, which were subsequently verified through expert validation.

Data Curation: The data curation pipeline consists of a series of cascaded operations. We filtered out samples containing hyperlinks, excessive text length with more than 4096 words, and zero answers; these steps eliminated 1,965,220, 5,310, and 734,011 instances, respectively. To focus on users exhibiting personalized preferences, we selected only those who posted at least three CIPD questions across all domains. From an initial pool of 6.8 million StackExchange questions, the final dataset contains 626,786 questions grouped by 15,963 users.

Notations: define our dataset as $\mathcal{U} = \{U_1, U_2, \ldots, U_N\}$, where N represents the total number of users ($N = 15,963$). Each U_i is associated with a sequence of data instances $(u_{i,1}, u_{i,2}, \ldots, u_{i,T_i})$, where T_i denotes the number of tracked QA pairs for user i. The instance $u_{i,j}$ for any $j \in \{1, 2, \cdots, T_i\}$ is defined as a tuple:

$$u_{i,j} = \left(q_{i,j}, (a^*_{i,j}, v^*_{i,j}), \left\{ (a^{(k)}_{i,j}, v^{(k)}_{i,j}) \right\}_{k=1}^{N_{i,j}} \right). \quad (1)$$

In Eqn. (1), $q_{i,j}$ denotes the j-th question of user i. The pair $(a^*_{i,j}, v^*_{i,j})$ represents the user-selected (accepted) answer and its corresponding vote count. Similarly, $(a^{(k)}_{i,j}, v^{(k)}_{i,j})$ represents the k-th candidate answer along with its vote count, where $N_{i,j}$ denotes the total number of collected answers for question $q_{i,j}$. We define $(a^t_{i,j}, v^t_{i,j})$ as the answer with the maximal vote count and that specific vote count, respectively. In our CIPD dataset, $a^t_{i,j} \neq a^*_{i,j}$ and $v^t_{i,j} > v^*_{i,j}$.

Rationale extraction: We employed an LLM to extract personalized rationales from the QA trajectory of user i. To ensure computational efficiency, we input each instance $u_{i,j}$, comprising the current question $(q_{i,j})$, the explicitly marked user-selected answer (a^*), and the remaining answers $a^{(k)}_{i,j}$ for any $k \in [N_{i,j}]$, along with a context of at most recent 50 historical questions $(q_{i,j-49}, \cdots, q_{i,j})$, into the LLM, which was then prompted to deduce the plausible cognitive drives behind the user's choices. Specifically, we enforced a structured JSON output schema that steers the generation through explicit reasoning steps: identifying a behavioral label ("reason"), mapping it to a descriptive psychological framework ("theory"), and synthesizing a

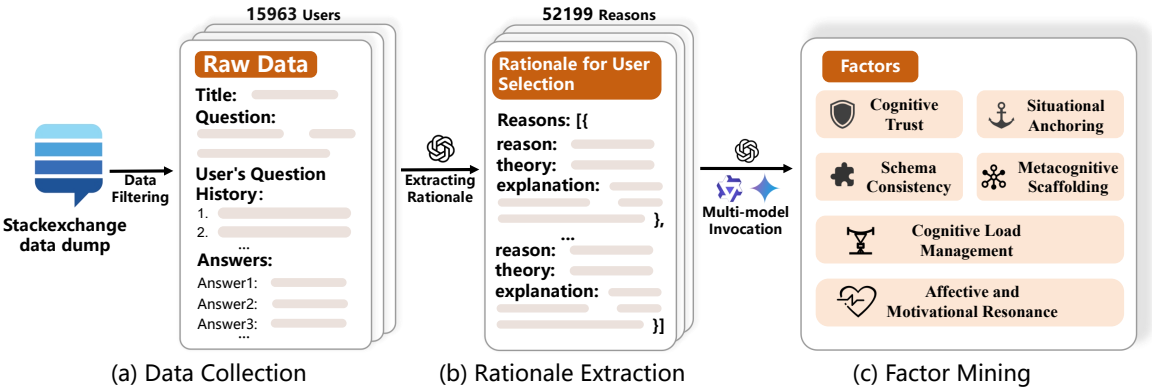


Figure 3: The pipeline of the Factor Distillation process. We first collect and filter raw CIPD data (a), then employ LLMs to extract decision rationales behind user choices (b), and finally distill these rationales into six cognitive factors through multi-model consensus (c).

justification for this alignment ("explanation"). To validate our approach, we randomly sampled 500 instances and conducted a manual inspection to assess the reliability of the generated explanations. The specific evaluation criteria are detailed in Appendix B.1. We invited two domain experts to perform this evaluation, obtaining a Cohen's Kappa of 0.7251 and a 97% alignment rate with expert judgment, which together support the reliability of our extraction pipeline.

Factor Distilling: Building upon the aforementioned relationships, we aggregated the discrete rationales into a candidate pool and employed LLMs to distill core influencing factors. The prompt is shown in Figure A2 of Appendix B. To ensure the stability and robustness of the extraction results, we adopted an ensemble strategy involving multiple models and repeated invocations, selecting factors with the highest consensus via frequency statistics. Subsequently, to conduct a qualitative plausibility assessment, two expert volunteers with Ph. D.s in Psychology were invited to manually evaluate and verify the results. The six distinct factors identified through this process are: (1) **Cognitive Trust** (CT): Does the response align with the user's threshold for trust and credibility within this specific domain? (2) **Situational Anchoring** (SA): Is the response precisely calibrated to the user's immediate context and specific problem? (3) **Schema Consistency** (SC): Does the response integrate coherently with the user's prior knowledge and existing mental models? (4) **Cognitive Load Management** (CLM): Is the complexity of the response tailored to match the user's cognitive capacity? (5) **Metacognitive Scaffolding** (MS): Does the response provide structural support that fosters the user's critical thinking skills? (6) **Affective and Motivational Resonance** (AMR): Does the response resonate with the user's current emotional state and motivational orientation? Three specific cases and the expert evaluation criteria are provided in Appendix B.2.

4 The CoPA Benchmark

We establish the CoPA benchmark to rigorously evaluate personalization methods using the six core factors. These factors serve as fine-grained dimensions for assessing the alignment between model-generated content and user cognitive needs.

4.1 Benchmark Construction

Data collection. We constructed the evaluation set by sampling the top 20% of users per domain, ranked by the vote counts of their most recent questions, to ensure the data quality. Domains with fewer than 10 users were fully included, while domains with more than 50 candidates were restricted to 50 users. This process yielded 1,985 unique users. Finally, these users and their corresponding interaction histories constitute our CoPA benchmark.

User Profile Construction. For each user i for any $i \in [N]$ in the benchmark, we constructed the user preference profile, denoted as P_i, across the six identified factors. By feeding an LLM with the user's 50 recent interactions, we prompted the model to quantify the user's preference for each factor $f_k \in \{\text{CT, SA, SC, CLM, MS, AMR}\}$. This process yielded a structured profile $P_i = \{p_1, p_2, ..., p_6\}$, where p_k represents the user's preference on factor f_k.

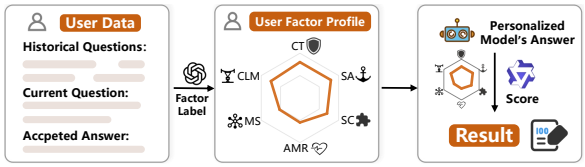


Figure 4: The proposed evaluation pipeline.

Statistics	Eng. Tools	Sci. Theory	Life. Society	Leisure Fandom
Users	864	456	413	252
Records (Avg. Q/User)	30.69	27.57	41.09	40.74
Q. Title Length	8.96	9.92	9.64	9.69
Q. Body Length	128.21	131.23	109.49	135.85
Factors Profile Size	555.12	597.74	580.34	590.10

Table 2: Dataset statistics of the CoPA benchmark.

Table 2 presents the detailed statistics of the CoPA benchmark constructed via this pipeline.

Evaluation Metric. We evaluate the response R_i generated for user i by measuring its alignment with the established profile P_i. An evaluator LLM compares R_i against each $p_k \in P_i$ to assign an alignment score s_k for any $k \in \{1, 2, \cdots, 6\}$. Prompt details are left in Appendix C, Figure A4. We adopt a 3-point Likert scale, as commonly used in recent LLM-based evaluation studies (Baek et al., 2024; Salemi and Zamani, 2025), to quantify the alignment score s_k:

- 0 (Mismatch): The response violates or ignores user preferences.
- 1 (Partial Match): The response aligns partially but lacks sufficient depth.
- 2 (Full Match): The response perfectly adapts to the user's factor profile.

The final scores are normalized to the range $[0, 1]$ for aggregation. Figure 4 demonstrates the user profile construction and evaluation process.

4.2 Evaluated Approaches

We evaluated four representative approaches for personalized QA on our benchmark.

No-Personalization generates responses directly using the LLM without access to user profiles. It constitutes a general-purpose, non-personalized baseline. The prompt is shown in Figure A9 of Appendix G.

Time-Personalization (Cattell, 1963) assumes that a user's cognitive state evolves over time. We construct the user profile using the K most recent historical questions in chronological order to prompt the LLM. The prompt is in Figure A10.

RAG-Personalization (Salemi et al., 2024b) retrieves the top-K historical questions with the highest semantic affinity to the current query. By aggregating these relevant instances, the model infers the user's specific attentional focus to generate tailored responses. The prompt is shown in Figure A11 of Appendix G.

Profile-Personalization (Su et al., 2025) compresses the user's sequence of K recent questions into Domain Profiles and Global Profiles, thereby extracting higher-dimensional personalized information. See Appendix G.3 for details.

5 Experiments

5.1 Experimental Setup

For the rationale extraction task, we employ the GPT-5-chat-latest model with the temperature parameter set to 0.7. In the Factor Mining phase, we ensembled the results from GPT-5-chat-latest[2], Gemini-2.5-Pro[3], and Qwen3-Max[4], all configured with a temperature of 0.7. Regarding the four approaches for personalized QA, we select the Qwen3-8B (Fast-Thinking mode) (Yang et al., 2025) and the GPT-4o-mini (Hurst et al., 2024) to serve as backbone models. Furthermore, within the RAG personalization module, Qwen3-Embedding-4B (Zhang et al., 2025b) is adopted as the embedding model. For evaluation, we use Qwen3-32B (Fast-Thinking mode) (Yang et al., 2025). In terms of hardware and deployment, all local models are hosted on a computing platform equipped with four NVIDIA A800 GPUs, utilizing the vLLM[5] framework for inference acceleration.

5.2 Effectiveness of Factors

We conducted a quantitative evaluation on CoPA to validate the alignment of our factor-based scoring with ground-truth user preferences. We utilized the factor profile constructed from the preceding CIPD question to evaluate the accepted and top-voted answers, obtaining scores denoted as S_{acc} and S_{top}, respectively. We defined three evaluation metrics: **Accuracy**: The proportion of instances where $S_{acc} > S_{top}$. **Tie Rate**: The proportion of instances where $S_{acc} = S_{top}$. **Score Margin**: The difference between S_{acc} and S_{top}, calculated as $S_{acc} - S_{top}$. An evaluation metric that effec-

[2] https://chatgpt.com/
[3] https://gemini.google.com/
[4] https://chat.qwen.ai/
[5] https://docs.vllm.ai/en/stable/

tively captures user preference should maximize Accuracy and Margin while minimizing the Tie Rate, thereby demonstrating a stronger capability to distinguish the user's personal choice from the community consensus. We compared the factor-based scoring with four evaluation methods: (1) Direct scoring via LLM-as-a-Judge on a 3-point Likert scale; (2) CoT scoring via LLM-as-a-Judge with few-shot demonstrations and explicit reasoning steps on a 3-point Likert scale; (3) heuristic-based evaluation metrics (Jaccard and Inclusion Coefficient) derived from (Su et al., 2025) (these two metrics are not normalized; therefore, the Tie Rate and Score Margin are not calculated.); (4) scoring using randomly generated factors by GPT-5-chat-latest, adapted from the methodology in (Salemi et al., 2024b). Crucially, to prevent data leakage, all data samples in CoPA benchmark were truncated to the preceding CIPD question for factor construction. The prompts for the Direct, CoT, and Random baselines are provided in Appendix D.

The results are presented in Table 3. Distinguishing "Accepted" from "Top-voted" answers proves to be a non-trivial challenge. Top-voted answers often feature high quality, serving as potent distractors for the evaluator LLM. This is evident in the Direct baseline, which exhibits excessive Tie Rates and low Accuracy. Although CoT scoring improves over Direct scoring, it still leaves Tie Rates above 55% in all domains and Accuracy below 32%, indicating that more elaborate prompting alone is insufficient. Compared to Direct and CoT scoring, our evaluation based on distilled factors consistently reduces Tie Rates to 18.24%–26.28% and improves Accuracy to 51.43%–55.37%, demonstrating that distilled factors enable the identification of granular features aligned with user intent. Relative to heuristic metrics (Jaccard/Inclusion), which perform better than Direct scoring but rely on surface-level lexical overlap, our method maintains a clear lead by capturing deeper cognitive alignment. Against the baseline with randomly generated factors, distilled factors still yield sizable gains across all domains, confirming the effectiveness of data-driven factor distilling rather than prompt sophistication alone. For completeness, the corresponding results when the factor profile is constructed from the current CIPD question are reported in Appendix D, Table A4.

To investigate the effectiveness of each factor, we conducted an ablation study (Table 4). The results indicate that removing any individual factor leads to a degradation in performance in terms of Accuracy, Tie Rate, and Score Margin, thereby demonstrating the necessity of each factor in capturing the user preferences.

5.3 Generalizability of the Factors

To validate the generalizability of the mined factors, we extended our evaluation to two external benchmark datasets: UPGC-QA (Su et al., 2025) and LaMP-QA (Salemi and Zamani, 2025). We sought to evaluate whether factors can enhance the performance of baselines on these benchmarks. Specifically, we integrated our factors into the system prompts of the baseline methods, guiding the models to focus on these dimensions during generation. As shown in Table A5 and Table 5 (the prompt and result are detailed in Appendix E.), incorporating these factors, yielded consistent and significant performance gains across two distinct benchmark, two different backbone models, and all evaluated personalization methods. This improvement indicates the factors' generalizability, suggesting that their optimization is a promising direction for personalized QA and thereby motivating the use of our factor-based benchmark, CoPA.

5.4 Factor-level evaluation of personalized QA approaches using CoPA

Table 6 presents a comparative evaluation of Time-, RAG-, and Profile-Personalization methods against the non-personalized baseline and ground truth across four categories.

The results demonstrated that incorporating personalized context consistently yields performance gains in general. Among the three personalization strategies, Profile-Personalization demonstrated the most superior performance on all four subdomains. This suggests that constructing structured and semantically rich user profiles allows for a more precise capture of user intent, thereby generating responses that are better aligned with user needs. Additionally, the performances of RAG-Personalization and Time-Personalization surpassed the top-voted answer, indicating that leveraging retrieval augmentation and temporal information serves as an effective personalization mechanism.

Regarding the comparison of base models, Qwen 3 8B outperformed GPT-4o-mini overall on this benchmark. Even in the No-Personalization setting, the average score of Qwen 3 8B surpassed

Method	Eng. & Tools			Science			Lifestyle			Leisure		
	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑
Direct	12.15	80.67	0.026	18.86	75.00	0.065	10.17	84.75	0.027	11.90	81.75	0.028
CoT	26.00	60.31	0.061	31.35	55.16	0.091	26.39	61.74	0.079	29.55	59.85	0.106
Jaccard	40.28	–	–	42.32	–	–	39.23	–	–	42.06	–	–
Incl. Coeff.	34.07	–	–	26.54	–	–	27.60	–	–	28.17	–	–
Random	25.58	61.00	0.049	30.26	56.14	0.074	33.90	53.27	0.081	35.09	50.00	0.071
ours	**51.57**	**26.28**	**0.127**	**53.44**	**22.88**	**0.136**	**51.43**	**24.12**	**0.093**	**55.37**	**18.24**	**0.143**

Table 3: Comparison of Evaluation Methodologies. Arrows (↑/↓) indicate whether higher or lower values are better. Values are in percentages (%) except for Margin.

Factors	Eng. & Tools			Science			Lifestyle			Leisure		
	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑
w/o CT	54.28	29.40	0.177	54.82	31.36	0.203	54.96	26.15	0.172	54.76	26.19	0.198
w/o SA	54.05	31.02	0.182	52.41	33.99	0.208	55.21	25.91	0.173	55.95	26.59	0.208
w/o SC	54.63	29.17	0.181	54.17	32.02	0.206	54.72	26.39	0.174	56.35	26.19	0.207
w/o CLM	53.24	31.37	0.175	53.95	32.68	0.206	55.69	24.94	0.171	57.14	26.19	0.216
w/o MS	50.00	34.61	0.174	51.54	35.96	0.199	53.51	28.57	0.170	53.57	28.17	0.200
w/o AM	53.82	30.21	0.181	53.73	33.11	0.203	55.21	26.88	0.174	55.56	26.19	0.208
All	**55.09**	**28.24**	**0.189**	**55.04**	**30.92**	**0.211**	**56.17**	**24.21**	**0.180**	**57.14**	**24.21**	**0.216**

Table 4: Ablation Study of Different Factors. Arrows (↑/↓) indicate whether higher or lower values are better. Values are in percentages (%) except for Margin.

Method	Wildchat		StackEx.		CS101	
	Jac.	I.C.	Jac.	I.C.	Jac.	I.C.
Qwen3 8B						
DirectQA	6.52	8.29	40.64	64.86	15.56	42.86
RAGQA	6.57	8.64	40.72	65.22	16.59	43.02
w/ Factors	**7.88**	**9.33**	**41.99**	**66.71**	**17.15**	**43.96**
ProfileQA	6.99	8.87	43.01	73.02	16.32	45.99
w/ Factors	**8.31**	**10.03**	**44.17**	**74.23**	**17.67**	**47.35**
GPT-4o-mini						
DirectQA	7.11	9.27	42.33	64.79	15.77	43.13
RAGQA	7.86	9.89	43.34	65.23	16.94	43.21
w/ Factors	**8.31**	**10.22**	**44.57**	**67.14**	**17.11**	**43.99**
ProfileQA	8.22	11.35	44.09	72.98	16.37	46.23
w/ Factors	**9.64**	**11.67**	**44.96**	**75.05**	**18.24**	**47.98**

Table 5: Comparison of personalization methods with and without factor augmentation on three datasets. **Jac.**: Jaccard, **I.C.**: Incl. Coeff.

that of GPT-4o-mini. When combined with Profile-Personalization, Qwen 3 8B further narrowed the gap with human-accepted answers, demonstrating the model's strong potential in handling personalized question-answering tasks.

From a category perspective, models generally excelled in the Engineering & Tools but underperformed in Science & Theory. We attribute this disparity to the latter's inherent demand for deep reasoning and specialized domain knowledge, which challenges current personalization models. This finding highlights a critical avenue for future research: enhancing the capability of personalized systems to navigate complex theoretical and high-cognitive-load scenarios.

Our detailed factor-level analysis (Appendix G.1) reveals several key insights. The inclusion of user profiles yielded substantial gains across Cognitive Trust, Schema Consistency, and Situational Anchoring. This indicates that explicit user modeling enhances contextual alignment and fosters deeper cognitive trust. In stark contrast, Metacognitive Scaffolding remains a significant challenge, with all models scoring exceptionally low, highlighting the inherent difficulty of inducing user reflection in generative tasks. Conversely, Cognitive Load Management scores were consistently high and were further improved by personalization, demonstrating efficacy in tailoring information density. Finally, marked improvements in Affective and Motivational Resonance confirm that profile-aware generation moves beyond mere information delivery to achieve genuine user resonance by capturing

Method	E&T	S&T	L&S	L&F	Avg.
Accepted Ans.	**0.8009**	**0.7984**	**0.7393**	**0.7946**	**0.7833**
Random Ans.	0.6183	0.5774	0.5516	0.5418	0.5722
Top Ans.	0.6207	0.5941	0.5633	0.5886	0.5917
GPT-4o-mini					
No-Pers.	0.6307	0.4700	0.5347	0.5136	0.5373
Time-Pers.	0.6858	0.5353	0.5591	0.5516	0.5830
RAG-Pers.	0.6964	0.5334	0.5748	0.5472	0.5909
Profile-Pers.	**0.7003**	**0.5912**	**0.6223**	**0.5719**	**0.6214**
Qwen 3 8B					
No-Pers.	0.6671	0.5409	0.5803	0.5403	0.5821
Time-Pers.	0.7188	0.5691	0.6230	0.5638	0.6187
RAG-Pers.	0.7198	0.5724	0.6172	0.5721	0.6204
Profile-Pers.	**0.7341**	**0.6365**	**0.6729**	**0.6002**	**0.6609**

Table 6: Performance comparison including gold standard references and different personalization models. **E&T**: Engineering & Tools, **S&T**: Science & Theory, **L&S**: Lifestyle & Society, **L&F**: Leisure & Fandom.

Factor	Human-Human (κ)	Human-LLM (ρ)
CT	0.622	0.733
SA	0.687	0.830
SC	0.609	0.744
CLM	0.641	0.744
MS	0.663	0.683
AMR	0.637	0.788
Overall	0.643	0.754

Table 7: Human-Human reports the agreement between the two annotators measured by Cohen's weighted Kappa, and Human-LLM reports the Spearman correlation between the average human ratings and the LLM scores.

emotional nuances.

5.5 Human Validation of CoPA Scoring

To assess whether CoPA scoring is reasonably aligned with human judgments, we recruited two graduate annotators to independently evaluate 50 randomly sampled question-response pairs using the same 3-point Likert rubric and six-factor criteria as our LLM evaluator. Detailed instructions are provided in Appendix F. As shown in Table 7, the annotators achieved an overall weighted Cohen's Kappa of 0.643, indicating substantial inter-rater agreement. Moreover, the average human ratings showed a Spearman correlation of 0.754 with the LLM scores, suggesting that the proposed scoring scheme is reasonably consistent with human judgments.

5.6 Factor Correlation Analysis

To evaluate the relevance of our factors, We analyzed the inter-factor relationships using the Spearman correlation coefficient (Spearman, 1961), as shown in Figure 5. The results revealed moderate positive correlations across most factors, indicating that while they collectively contribute to personalization, they capture distinct information without significant redundancy. Specifically, the peak correlation between Cognitive Trust (CT) and Schema Consistency (SC) highlights a tight dependency: trust relies heavily on alignment with user schemas. However, the coefficient remains below the multicollinearity threshold ($r > 0.9$), confirming their *discriminant validity*—SC represents the establishment *pathway*, while CT reflects the resultant *state*. Conversely, Cognitive Load Management (CLM) displayed the lowest average correlation, underscoring its statistical independence. This confirms that high trust or resonance does not inherently imply low cognitive load. Thus, CLM captures a unique dimension of "information processing cost," validating its necessity as a distinct metric in the CoPA benchmark.

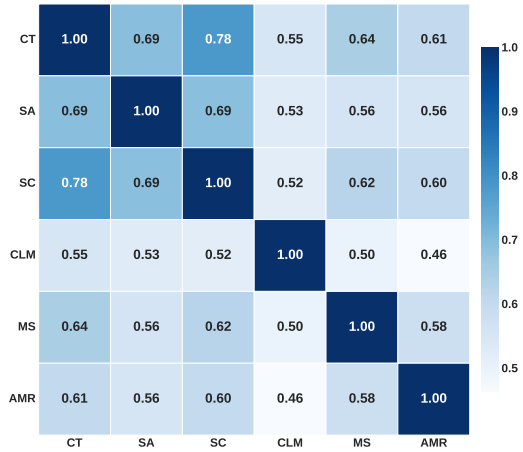


Figure 5: Matrix of Spearman correlations between the factors.

6 Related Work

Personalized Large Language Models. Personalization is critical in search (Baek et al., 2024; Sharma et al., 2024), recommendation (Wu et al., 2024; Liu et al., 2025c), and text generation (Li et al., 2024; Liu et al., 2025a). Existing personalized LLMs fall into two paradigms. The first, non-parametric context augmentation, leverages In-Context Learning to inject user history as external

knowledge. Techniques like Retrieval-Augmented Generation (Salemi et al., 2024a; Zhuang et al., 2024) retrieve relevant behavioral fragments to implicitly mimic user style, while others distill interaction data into explicit User Profiles (Chen et al., 2025; Su et al., 2025) to guide generation. The second paradigm, parametric modeling, internalizes personalized data. While traditional fine-tuning (Liu et al., 2025b; Tan et al., 2024) is effective, it suffers from high storage costs and catastrophic forgetting. Consequently, recent lightweight strategies (Zhang et al., 2025a; He et al.) propose training-free, dynamic construction of user-specific style matrices at inference time to address these bottlenecks.

Evaluation of Personalized Generation. Evaluating alignment with user preferences involves three main paradigms. First, traditional n-gram metrics like BLEU and ROUGE (Salemi et al., 2024b; Kumar et al., 2024) measure lexical overlap but struggle with the open-ended, semantic nature of personalized contexts. Second, the mainstream LLM-as-a-Judge approach (Salemi and Zamani, 2025; Dong et al., 2024) utilizes LLMs to assess semantic dimensions (e.g., helpfulness) in a reference-free setting. Third, rule-based heuristics (Su et al., 2025; Dai et al., 2024) employ customized measurement standards. However, current metric often lack empirical grounding and fail to elucidate the cognitive determinants driving personalized decisions, highlighting the need for data-driven quantitative standards.

7 Conclusion

This paper investigates the Community-Individual Preference Divergence (CIPD) phenomenon within the StackExchange dataset. Through fine-grained user attribution analysis on extensive interaction data, we distilled six core factors. Leveraging these factors, we introduce CoPA, a benchmark designed to evaluate Large Language Models in personalized question answering. Systematic experiments with established baselines demonstrated that our factor-based metrics effectively quantify personalization. Furthermore, empirical results showed that incorporating user profiles significantly enhanced both response quality and preference alignment.

Acknowledgments

This work was supported by Zhongguancun Academy (Grant Nos. C20250513 and XTS0025) and the National Natural Science Foundation of China (Grant Nos. 62277017 and 62137001).

Limitations

This study has several limitations.

First, our work distills six core factors based on the StackExchange dataset. A potential limitation is that these factors may exhibit interdependencies, and their complete orthogonality cannot be guaranteed. Furthermore, the reliance on a single data source might restrict the comprehensiveness of the identified factors. Given the dataset's strong pedagogical and explanatory orientation, the distilled factors may over-represent explanatory preferences prevalent in learning-centric communities. This, in turn, could limit the generalizability of our framework to other personalization scenarios, such as in conversational agents or recommender systems. Future work should therefore aim to integrate heterogeneous data sources to enhance the completeness of factor discovery and to evaluate and adapt our framework across a broader spectrum of application domains.

Second, our framework heavily relies on Large Language Models (LLMs) at multiple stages, including principle induction, factor distillation, and evaluation. Although we employed strategies such as context truncation and controlled prompting to ensure the quality of the generated outputs, this dependency may introduce epistemic circularity and inherent model biases. A crucial direction for improving the framework's robustness is to incorporate a human-in-the-loop validation mechanism.

Third, the stability and reliability of the current "LLM-as-a-Judge" mechanism are susceptible to the intrinsic biases of LLMs. Consequently, future research will prioritize two avenues: 1) developing more sophisticated and fine-grained Judge Models, and 2) operationalizing the abstract theoretical factors into concrete, measurable metrics under a more rigorous theoretical grounding to enhance the objectivity and accuracy of the evaluation.

Ethics Statement

Potential ethical concerns arise from the use of datasets collected from open-source web platforms, which may carry the risk of privacy leakage. To mitigate this, we have manually sanitized the datasets by redacting all Personally Identifiable Information, such as usernames and identifiers. Furthermore, we acknowledge that content generated

by Large Language Models (LLMs) is susceptible to hallucinations and ethical pitfalls. Consequently, a rigorous evaluation of the reliability and safety of the model outputs is essential.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674.

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen Herring, and Sujay Kumar Jauhar. 2024. Knowledge-augmented large language models for personalized contextual query suggestion. In *Proceedings of the ACM Web Conference 2024*, pages 3355–3366.

Raymond B Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1.

Aili Chen, Chengyu Du, Jiangjie Chen, Jinghan Xu, Yikai Zhang, Siyu Yuan, Zulong Chen, Liangyue Li, and Yanghua Xiao. 2025. DEEPER insight into your user: Directed persona refinement for dynamic persona modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24157–24180, Vienna, Austria. Association for Computational Linguistics.

Zhenlong Dai, Chang Yao, Wenkang Han, Ying Yuan, Zhipeng Gao, and Jingyuan Chen. 2024. MPCoder: Multi-user personalized code generator with explicit and implicit style representation learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3765–3780, Bangkok, Thailand. Association for Computational Linguistics.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L Schrum, and Anca Dragan. Context steering: Controllable personalization at inference time. In *The Thirteenth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11196–11215.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024*, pages 3367–3378.

Ben Liu, Jihai Zhang, Fangquan Lin, Xu Jia, and Min Peng. 2025a. One size doesn't fit all: A personalized conversational tutoring agent for mathematics instruction. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2401–2410.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025b. LLMs + persona-plug = personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9373–9385, Vienna, Austria. Association for Computational Linguistics.

Langming Liu, Shilei Liu, Yujin Yuan, Yizhen Zhang, Bencheng Yan, Zhiyuan Zeng, Zihao Wang, Jiaqi Liu, Di Wang, Wenbo Su, et al. 2025c. Uqabench: Evaluating user embedding for prompting llms in personalized question answering. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5652–5661.

Kevin Kyung Nam, Mark S Ackerman, and Lada A Adamic. 2009. Questions in, knowledge in? a study of naver's question answering community. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 779–788.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large

language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 752–762.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392.

Alireza Salemi and Hamed Zamani. 2025. LaMP-QA: A benchmark for personalized long-form question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1159, Suzhou, China. Association for Computational Linguistics.

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 2025. Retrieval augmented generation with collaborative filtering for personalized text generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1294–1304.

Harmanpreet Singh, Nikhil Verma, Yixiao Wang, Manasa Bharadwaj, Homa Fashandi, Kevin Ferreira, and Chul Lee. 2024. Personal large language model agents: A case study on tailored travel planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 486–514, Miami, Florida, US. Association for Computational Linguistics.

Charles Spearman. 1961. The proof and measurement of association between two things.

Hang Su, Yun Yang, Tianyang Liu, Xin Liu, Peng Pu, and Xuesong Lu. 2025. Personalized question answering with user profile generation and compression. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4744–4763, Suzhou, China. Association for Computational Linguistics.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd international conference on World wide web*, pages 515–526.

Junda Wu, Cheng-Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian McAuley. 2024. Coral: collaborative retrieval-augmented large language models improve long-tail recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3391–3401.

Derong Xu, Xinhang Li, Ziheng Zhang, Zhenxi Lin, Zhihong Zhu, Zhi Zheng, Xian Wu, Xiangyu Zhao, Tong Xu, and Enhong Chen. 2025. Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25570–25578.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. Hydra: Model factorization framework for black-box llm personalization. *Advances in Neural Information Processing Systems*, 37:100783–100815.

A Question Distribution

Figure A18 shows the distribution of questions across various domains in the Engineering&Tools category. Figure A19 shows the distribution of questions across various domains in the Science&Theory category. Figure A20 shows the distribution of questions across various domains in the Lifestyle&Society category. Figure A21 shows the distribution of questions across various domains in the Leisure&Fandom category.

B Factor Mining

Figure A1 shows the prompt used for Rationale Extraction. Figure A2 shows the prompt used for Factor Mining.

B.1 Manual Inspection Criteria for Reason Analysis

We evaluate the Rationale Extraction based on the following criteria:

1 - Pass if the rationale is both logically coherent and theoretically sound. This requires the model to accurately identify an underlying motive and cite a relevant theory (e.g., "Social Identity Theory," "Cognitive Dissonance") that strongly supports it. The output must contain no factual errors.

0 - Fail if the rationale exhibits any of the following issues:

1. **Hallucination or Contradiction:** The inferred motive is irrelevant to, or contradicts, the user's provided conversation history.
2. **Theoretical Error:** The model fabricates a non-existent theory or concept.
3. **Inappropriate Application of Theory:** A valid theory is cited, but it is completely mismatched with the inferred motive (e.g., applying "Economic Utility Theory" to explain a rationale for "emotional counseling").

B.2 Factors and Their Associated Causal Cases

TableA1, A2, and A3 show three examples of factors associated with the reasons.

We evaluate the factors based on the following criteria:

1. Whether the factors are subject to cognitive biases.

System:
You are an experienced educationalist who has researched in the field of education for many years. Your specialty is to analyze a user's psychology and choices from a pedagogical perspective to understand their learning behavior and needs.

User:
User's historical question:
{Historical_questions}
User's current question:
{Current_question}
The answer to the current question:
{Answers}
The 'IsAccept' field in the answer data signifies user acceptance. Your task is to analyze the reasons why the user accepted this specific answer over other alternatives that had higher scores.

The response should be in JSON format:

```
{
  "reasons": [
    {
      "reason": "",
      "theory":"",
      "explanation": ""
    },
    {
      "reason": "",
      "theory": "",
      "explanation": ""
    }
  ],
  "global_explanation": ""
}
```

Requirements:
1. Each reason should be concise and not exceed five words.
2. Each reason should be grounded in an educational or psychological framework, incorporating appropriate academic terminology.
3. There is no limit to the number of reasons.
4. The theory embodies the educational principles reflected in this reason.
5. The 'global_explanation' should provide an overall summary of the reasons listed above.

Figure A1: Extract rationale Prompt.

System:
You are a researcher in the field of education with expertise in pedagogy, psychology, and linguistics.

User:
The pool of reasons for users' choices.
{Reasons}
The pool of reasons contains the rationales for why different users selected their answers. Your task is to analyze these rationales from an educational perspective to derive a set of quantifiable factors.

The response should be in JSON format:

```
{
  "factors": [
    {
      "factor": "",
      "example": "",
      "explanation": ""
    },
    {
      "factor": "",
      "example": "",
      "explanation": ""
    }
  ],
  "global_explanation": ""
}
```

Requirements:
1. You can use academic terminology to describe these factors.
2. The example section illustrates the conflicting reasons you found. It must contain at least five pairs of contrasting examples, with one supporting a choice (a positive example) and one opposing it (a negative example).
3. There is no limit to the number of factors.
4. The 'global_explanation' should provide an overall summary of the factor listed above.

Figure A2: Factor Mine Prompt.

2. Whether these factors constitute dimensions that influence an individual's decision-making process.

3. Whether the factors are generalizable across different populations, cultures, and contexts.

C Factor Label

Figure A3 shows the prompt used for factor label. Figure A4 shows the prompt used for evaluation .

D Effectiveness of Factors

Figures A5, A6, and A7 show the prompts used for the Direct, CoT, and Random baselines, respectively. Table A4 reports the main experimental results when the factor profile for evaluation is constructed from the current CIPD question.

E Generalizability of the Factors

Table5 show the result of UPGC-QA. we select the open-source Qwen3-8B (Fast-Thinking mode) (Yang et al., 2025) and the proprietary GPT-4o-mini (Hurst et al., 2024) to serve as backbone models. Figure A8 shows the factor prompt that is integrated into the system prompt. We adopt the experimental settings and evaluation metrics from the original papers to assess the effectiveness of our proposed factors.

F Instructions for Human Annotators

You will be given three pieces of information: (1) a user factor profile, (2) a question, and (3) a candidate response. Your task is to judge how well the response matches the user's preferences on the following six dimensions: Cognitive Trust, Situational Anchoring, Schema Consistency, Cognitive Load Management, Metacognitive Scaffolding, and Affective and Motivational Resonance.

For each dimension, assign one score using a 3-point Likert scale:

- 0 (*Mismatch*): the response clearly violates or ignores the user's preference;
- 1 (*Partial Match*): the response partially addresses the user's preference, but lacks sufficient depth or completeness;
- 2 (*Full Match*): the response is well aligned with the user's preference on this dimension.

Please evaluate each sample independently based only on the provided materials. Do not assume any additional user background beyond the given profile. If you are uncertain between two scores, choose the more conservative one.

G Factor-level evaluation of personalized QA approaches using CoPA

Figure A9 shows the prompt used for No-Personalization. Figure A10 shows the prompt used for Time-Personalization. Figure A11 shows the prompt used for RAG-Personalization.

Table A1: Case1

Reason	Theory	Explanation
Narrative gap resolution	Cognitive closure	The answer filled a specific gap left by the anime adaptation, providing the user with a clear explanation from the manga, thus satisfying their need for a complete mental model of the story.
Source-authoritative detail	Epistemic trust	The response referenced information from the primary source material (manga), which the user perceives as more authoritative and trustworthy, increasing acceptance likelihood.
Canon consistency	Schema theory	The explanation aligned with the user's existing narrative schema of the series, integrating seamlessly without contradictions, which aids in knowledge assimilation.
Context-specific relevance	Situated learning	The answer directly addressed the specific context of the user's question (Touka's ability in Inari Konkon), creating immediate perceived practical relevance.
Global Explanation	The user accepted this answer because it directly addressed a specific narrative gap from the anime by referencing canonical manga evidence, which satisfied their cognitive need for closure. The information was consistent with their existing knowledge schema, delivered from what they perceived as an authoritative source, and situated within the precise narrative context they were curious about, all of which increased trust, relevance, and integration into their understanding.	
Factors	Cognitive Trust, Situational Anchoring, Schema Consistency	

G.1 Detailed factor-level analysis

Tables A6 –A9 show the detailed factor-level analysis.

G.2 Impact of User History Context Length (K) on Model Performance

Figure A12 illustrates the specific impact of incorporating varying lengths of user query history (K) on model performance across four sub-domains. Overall, the experimental results demonstrate a significant positive correlation between the inclusion of user history and model performance, with the full history (k = all) yielding the most substantial gains. However, it is important to note that increasing K inevitably leads to higher inference time costs. This trade-off provides valuable insights and directions for future research.

G.3 Profile-Personalization

Figures A13–A17 show the prompts used for Profile-Personalization.

H License

The code and benchmark resources for CoPA will be released under the Apache License 2.0. This license permits use, modification, and redistribution subject to the terms of the license.

I AI Assistance Usage

Gemini[6] was employed as a linguistic assistant to improve readability. Specifically, we used the tool to refine the preliminary drafts of certain sections. All AI-assisted text underwent rigorous manual review and editing to ensure precision and coherence before finalization.

[6] https://gemini.google.com/

Table A2: Case2

Reason	**Theory**	**Explanation**
Clear conceptual differentiation	Cognitive Load Theory	The answer helped reduce intrinsic cognitive load by clearly distinguishing EFHW and long wire antennas, which addressed the user's confusion and allowed them to process the technical differences more efficiently.
Direct application insight	Situated Learning Theory	The answer connected principles to the user's real-world scenario (hotel room limitations), enabling context-based understanding that made the information more relevant and actionable.
Mechanism-focused explanation	Constructivist Learning Theory	The answer explained how multiple loops would affect inductance, SWR curve, and radiation pattern, enabling the user to build mental models from prior knowledge.
Expectation management	Metacognition	By acknowledging the inevitable compromise in performance, the answer guided the user's realistic assessment of trade-offs, improving self-regulation in decision-making.
Global Explanation	The user accepted the answer because it reduced conceptual confusion between antenna types, contextualized the explanation to their space-limited operating environment, provided clear technical mechanisms, and managed expectations of performance. From a pedagogical perspective, the answer optimized cognitive load, fostered situated learning, supported constructivist model-building, and enhanced metacognitive awareness—leading to a response that was both technically informative and psychologically reassuring.	
Factors	Cognitive Load Management, Situational Anchoring, Schema Consistency, Metacognitive Scaffolding	

Table A3: Case3

Reason	Theory	Explanation
Growth mindset emphasis	Carol Dweck's Growth Mindset Theory	The accepted answer likely stressed that intelligence and skill can develop with effort, aligning with Dweck's research which shows children persist more when they view challenges as opportunities to grow rather than as fixed measures of ability.
Intrinsic motivation cultivation	Self-Determination Theory	By focusing on the satisfaction of mastering a difficult task, the answer may have encouraged autonomy and competence, supporting self-determination theory's emphasis on internal rewards over external pressures.
Scaffolded learning approach	Vygotsky's Zone of Proximal Development	The answer may have suggested breaking tasks into achievable steps, enabling the child to succeed incrementally while receiving guidance, which aligns with Vygotsky's principle of supporting learners just beyond their current capability.
Positive reinforcement use	Skinner's Operant Conditioning	Likely included recommendations of rewarding effort and persistence, teaching the child to associate trying again with positive outcomes, reinforcing the desired behavior.
Global Explanation	The user accepted the answer because it acknowledged the psychological roots of giving up after failure and addressed them through proven educational frameworks. By combining growth mindset principles, intrinsic motivation development, scaffolded support, and positive reinforcement, the answer offered a multi-dimensional strategy that resonates with both pedagogical and developmental psychology approaches, ensuring practical guidance aligned with the child's stage of cognitive and emotional growth.	
Factors	Metacognitive Scaffolding, Affective and Motivational Resonance, Schema Consistency	

Method	Eng. & Tools			Science			Lifestyle			Leisure		
	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑	Acc↑	Tie↓	Mar.↑
Direct	12.15	80.67	0.026	18.86	75.00	0.065	10.17	84.75	0.027	11.90	81.75	0.028
CoT	26.00	60.31	0.061	31.35	55.16	0.091	26.39	61.74	0.079	29.55	59.85	0.106
Jaccard	40.28	–	–	42.32	–	–	39.23	–	–	42.06	–	–
Incl. Coeff.	34.07	–	–	26.54	–	–	27.60	–	–	28.17	–	–
Random	46.18	44.00	0.184	50.22	41.67	0.210	47.94	40.68	0.169	52.78	35.32	0.233
ours	**55.09**	**28.24**	**0.189**	**55.04**	**30.92**	**0.211**	**56.17**	**24.21**	**0.180**	**57.14**	**24.21**	**0.216**

Table A4: Comparison of Evaluation Methodologies. Arrows (↑/↓) indicate whether higher or lower values are better. Values are in percentages (%) except for Margin.

System:
Role: You are an interdisciplinary domain expert. Your task is to evaluate the specific emphasis the user places on the following dimensions regarding the current question, based on an analysis of the user's historical query patterns and the response to the immediate question.

Evaluation Dimensions:
1. Cognitive Trust: What are the user's epistemic requirements regarding the credibility, reliability, and verifiability of the information?
2. Situational Anchoring: To what extent does the user require the response to be contextually aligned, practically applicable, or specific to a given scenario?
3. Schema Consistency: What is the nature of the user's existing prior knowledge and mental models (and how should the new information align with them)?
4. Cognitive Load Management: What are the user's constraints regarding information processing capacity and their tolerance for complexity?
5. Metacognitive Scaffolding: What are the user's requirements for structural guidance to facilitate higher-order understanding and self-regulated learning?
6. Affective and Motivational Resonance: What are the user's expectations regarding emotional engagement and motivational alignment within the response?

User:
User's Historical Questions:
{Historical_questions}
User's Current Question:
{Current_question}
The answer to the current question:
{Answer}

The response should be in JSON format:

```
{
   "Cognitive Trust": {
      "description": "", "explanation": ""
   },
   "Situational Anchoring": {
      "description": "", "explanation": ""
   },
   "Schema Consistency": {
      "description": "", "explanation": ""
   },
   "Cognitive Load Management": {
      "description": "", "explanation": ""
   },
   "Metacognitive Scaffolding": {
      "description": "", "explanation": ""
   },
   "Affective and Motivational Resonance": {
      "description": "", "explanation": ""
   }
}
```

Requirements:
1. Each description must be tailored to the user's specific circumstances.
2. Each description should provide a specific and accurate summary of the user's personal profile, with no word limit.
3. The explanation serves as the rationale for the description provided above.
4. Ensure the output adheres to the specified format.

Figure A3: Factor Label Prompt.

Method	A&E	L&P	S&C	Avg.
Qwen3 8B				
No-Pers.	0.3422	0.4824	0.4824	0.4357
RAG-Pers.	0.3634	0.4700	0.4983	0.4439
w/ Factors	**0.3743**	**0.4892**	**0.5131**	**0.4589**
GPT-4o-mini				
No-Pers.	0.3941	0.5162	0.5282	0.4795
RAG-Pers.	0.4182	0.4861	0.5343	0.4795
w/ Factors	**0.4378**	**0.5234**	**0.5720**	**0.5111**

Table A5: Performance comparison on LaMP-QA. **A&E**: Art & Entertainment, **L&P**: Lifestyle & Personal Development, **S&C**: Society & Culture. "w/ Factors" denotes the baseline augmented with our factor prompts.

Method	Cognitive Trust	Situational Anchoring	Schema Consistency	Cognitive Load Management	Metacognitive Scaffolding	Affective & Motiv. Resonance
GPT-4o-mini						
No-Personalization	0.5972	0.6348	0.6574	0.7737	0.4213	0.6997
Time-Personalization	0.6562	0.6985	0.7153	0.8183	0.4641	0.7622
RaG-Personalization	0.6719	0.7020	0.7338	0.8200	0.4797	0.7714
Profile-Personalization	0.6743	0.7101	0.7344	0.8163	0.4835	0.7832
Qwen3-8b						
No-Personalization	0.6441	0.6765	0.6973	0.7726	0.4589	0.7535
Time-Personalization	0.6933	0.7297	0.7564	0.8403	0.5000	0.7934
RaG-Personalization	0.6921	0.7367	0.7569	0.8299	0.5111	0.7922
Profile-Personalization	0.7135	0.7442	0.7841	0.8200	0.5272	0.8154

Table A6: Detailed factor-level analysis of Engineering&Tools

Method	Cognitive Trust	Situational Anchoring	Schema Consistency	Cognitive Load Management	Metacognitive Scaffolding	Affective & Motiv. Resonance
GPT-4o-mini						
No-Personalization	0.4452	0.5197	0.5022	0.5855	0.2621	0.5055
Time-Personalization	0.5044	0.5822	0.5680	0.6524	0.3279	0.5768
RaG-Personalization	0.5033	0.5789	0.5548	0.6579	0.3311	0.5746
Profile-Personalization	0.5513	0.6321	0.6285	0.7122	0.3982	0.6254
Qwen3-8b						
No-Personalization	0.5132	0.5888	0.5844	0.6568	0.3279	0.5746
Time-Personalization	0.5296	0.6283	0.6118	0.6941	0.3487	0.6020
RaG-Personalization	0.5340	0.6305	0.6173	0.6787	0.3596	0.6140
Profile-Personalization	0.6151	0.6974	0.6842	0.7303	0.4254	0.6667

Table A7: Detailed factor-level analysis of Science&Theory

Method	Cognitive Trust	Situational Anchoring	Schema Consistency	Cognitive Load Management	Metacognitive Scaffolding	Affective & Motiv. Resonance
GPT-4o-mini						
No-Personalization	0.4855	0.5896	0.5581	0.6961	0.2603	0.6186
Time-Personalization	0.4891	0.6235	0.5835	0.7276	0.2906	0.6404
RaG-Personalization	0.5024	0.6356	0.5920	0.7470	0.3063	0.6659
Profile-Personalization	0.5139	0.7018	0.6577	0.7827	0.3766	0.7014
Qwen3-8b						
No-Personalization	0.5182	0.6465	0.6029	0.7312	0.3220	0.6610
Time-Personalization	0.5835	0.6743	0.6501	0.7688	0.3644	0.6973
RaG-Personalization	0.5678	0.6780	0.6392	0.7676	0.3656	0.6852
Profile-Personalization	0.6235	0.7300	0.7094	0.7930	0.4262	0.7554

Table A8: Detailed factor-level analysis of Lifestyle&Society

Method	Cognitive Trust	Situational Anchoring	Schema Consistency	Cognitive Load Management	Metacognitive Scaffolding	Affective & Motiv. Resonance
GPT-4o-mini						
No-Personalization	0.4405	0.5357	0.5357	0.6845	0.3274	0.5575
Time-Personalization	0.4901	0.5873	0.5813	0.7044	0.3512	0.5952
RaG-Personalization	0.4802	0.5675	0.5833	0.6944	0.3611	0.5972
Profile-Personalization	0.5310	0.5922	0.6117	0.7053	0.3566	0.6348
Qwen3-8b						
No-Personalization	0.4841	0.5774	0.5635	0.6825	0.3492	0.5853
Time-Personalization	0.5020	0.6032	0.5972	0.7004	0.3710	0.6091
RaG-Personalization	0.5198	0.6052	0.6032	0.7143	0.3611	0.6290
Profile-Personalization	0.5496	0.6349	0.6369	0.7044	0.4127	0.6627

Table A9: Detailed factor-level analysis of Leisure&Fandom

System:
Role: You are a fair and insightful judge with exceptional reasoning and analytical abilities. Your task is to evaluate whether the response to the user's question aligns with the user's factor profile.

Evaluation Criteria (The 6 Dimensions):
1. Cognitive Trust: What are the user's epistemic requirements regarding the credibility, reliability, and verifiability of the information?
2. Situational Anchoring: To what extent does the user require the response to be contextually aligned, practically applicable, or specific to a given scenario?
3. Schema Consistency: What is the nature of the user's existing prior knowledge and mental models (and how should the new information align with them)?
4. Cognitive Load Management: What are the user's constraints regarding information processing capacity and their tolerance for complexity?
5. Metacognitive Scaffolding: What are the user's requirements for structural guidance to facilitate higher-order understanding and self-regulated learning?
6. Affective and Motivational Resonance: What are the user's expectations regarding emotional engagement and motivational alignment within the response?

Scoring Rubric (3-point Likert scale):

- 0 (Mismatch): The response actively violates the user's preference or completely ignores a high-priority requirement defined in the profile.
- 1 (Partial Match): The response addresses the requirement but lacks depth, or only partially aligns with the user's preference.
- 2 (Full Match): The response perfectly adapts to the user's constraints and preferences described in the profile.

User:
Input Data:
<user_factor_profile>
{factors_profile}
</user_factor_profile>

<question>
{question}
</question>

<response_to_evaluate>
{response}
</response_to_evaluate>

The response should be in JSON format:
```
{
   "Cognitive Trust": { "score": 0, "reasoning": "Brief explanation..." },
   "Situational Anchoring": { "score": 0, "reasoning": "Brief explanation..." },
   "Schema Consistency": { "score": 0, "reasoning": "Brief explanation..." },
   "Cognitive Load Management": { "score": 0, "reasoning": "Brief explanation..." },
   "Metacognitive Scaffolding": { "score": 0, "reasoning": "Brief explanation..." },
   "Affective and Motivational Resonance": { "score": 0, "reasoning": "Brief explanation..." }
}
```

Requirements:
1. Analyze the match between the <user_factor_profile> and <response_to_evaluate> for EACH factor.
2. Assign a score of 0, 1, or 2. **Important: The 'score' field must be a raw INTEGER type (int), do not use strings.** (e.g., output 1, not "1").
3. Provide a brief reasoning for your score.
4. Output the result in strict JSON format.

Figure A4: Evaluation Prompt.

System:
Role: You are a fair and insightful judge with exceptional reasoning and analytical abilities. Your task is to evaluate whether the response to the user's question aligns with the user's profile.

Scoring Rubric (3-point Likert scale):

- 0 (Mismatch): The response actively violates the user's preference or completely ignores a high-priority requirement defined in the profile.
- 1 (Partial Match): The response addresses the requirement but lacks depth, or only partially aligns with the user's preference.
- 2 (Full Match): The response perfectly adapts to the user's constraints and preferences described in the profile.

User:
Input Data:
<user_profile>
{user_profile}
</user_profile>

<question>
{question}
</question>

<response_to_evaluate>
{response}
</response_to_evaluate>

The response should be in JSON format:

```
{
   "score": 0,
   "reasoning": ""
}
```

Requirements:
1. Analyze the match between the <user_profile> and <response_to_evaluate>.
2. Assign a score of 0, 1, or 2. **Important: The 'score' field must be a raw INTEGER type (int), do not use strings.**
3. Provide a brief reasoning for your score.
4. Output the result in strict JSON format.

Figure A5: Direct Evaluation Prompt.

System:
Role: You are a fair and insightful judge with exceptional reasoning and analytical abilities. Your task is to evaluate whether the response to the user's question aligns with the user's profile.

Scoring Rubric (3-point Likert scale):

- 0 (Mismatch): The response actively violates the user's preference or completely ignores a high-priority requirement defined in the profile.
- 1 (Partial Match): The response addresses the requirement but lacks depth, or only partially aligns with the user's preference.
- 2 (Full Match): The response perfectly adapts to the user's constraints and preferences described in the profile.

Evaluation Instructions (Chain-of-Thought):
1. Step 1 – Profile Analysis: Identify the key preferences and requirements revealed by the user's historical questions.
2. Step 2 – Response Analysis: Examine what the response actually provides (e.g., depth, style, format, domain focus).
3. Step 3 – Alignment Assessment: Compare the user's profile against the response. Note specific matches or mismatches.
4. Step 4 – Score Decision: Based on the above analysis, assign a score of 0, 1, or 2.

Output Format:
Return a single JSON object with the following fields:
{
"chain_of_thought": "Your step-by-step reasoning following Steps 1–4",
"score": 0,
"reasoning": "One-sentence summary of the final judgement"
}
IMPORTANT: The 'score' field must be a raw INTEGER type (int), not a string.

Demonstration Examples:
1. **Example 1 – Score: 0 (Mismatch).**
<user_profile> 1. How does backpropagation work mathematically in deep neural networks? 2. What is the derivation of the attention mechanism in Transformers? 3. Can you explain the proof of the universal approximation theorem? 4. What are the theoretical guarantees of gradient descent convergence? 5. How is the VC dimension defined and why does it bound generalization error? </user_profile>
<question>What is a neural network?</question>
<response_to_evaluate> A neural network is a machine learning model loosely inspired by the brain. It consists of layers of nodes (neurons) that transform input data step by step. Popular frameworks include TensorFlow and PyTorch. </response_to_evaluate>
Expected output:
{
"chain_of_thought": "Step 1 - Profile Analysis: All historical questions demand mathematical derivations and theoretical proofs, indicating an expert user. Step 2 - Response Analysis: The response is a beginner-level intuitive overview with no mathematics. Step 3 - Alignment Assessment: Direct conflict — the user expects formal rigor; the response delivers none. Step 4 - Score Decision: Clear mismatch.",
"score": 0,
"reasoning": "The response provides a superficial overview that completely ignores the user's demonstrated need for rigorous mathematical depth."
}
2. **Example 2 – Score: 2 (Full Match).**
<user_profile> 1. I am new to Python, how do I print hello world? 2. What does the if-else statement do in Python? 3. How do I create a list in Python and add items to it? 4. What is a for loop and how do I use it? 5. Can you explain what a function is in Python with a simple example? </user_profile>
<question>How do I read a file in Python?</question>
<response_to_evaluate> To read a file in Python, use the built-in open() function: with open("myfile.txt", "r") as f: content = f.read() print(content) The with statement closes the file automatically. You can also read line by line using a for loop, which you already know! </response_to_evaluate>
Expected output:
{
"chain_of_thought": "Step 1 - Profile Analysis: All questions are beginner-level Python syntax questions; user needs simple language and concrete examples. Step 2 - Response Analysis: Clear code example, plain language, connects to prior knowledge of for loops. Step 3 - Alignment Assessment: Perfect match — accessible, concrete, no unnecessary complexity. Step 4 - Score Decision: Full match.",
"score": 2,
"reasoning": "The response perfectly adapts to the beginner profile with simple language, a practical example, and a connection to prior knowledge."
}
3. **Example 3 – Score: 1 (Partial Match).**
<user_profile> 1. What are the latest advancements in large language model alignment research? 2. How does RLHF compare to Constitutional AI in terms of safety properties? 3. What are the known failure modes of reward hacking in AI systems? 4. Can you explain the debate around instrumental convergence in AI safety? 5. What is the current state of interpretability research for transformer models? </user_profile>
<question>What are the risks of deploying large language models in production?</question>
<response_to_evaluate> Deploying LLMs in production carries risks such as hallucination, bias, high compute costs, and potential misuse. Implement guardrails, monitoring, and human oversight. </response_to_evaluate>
Expected output:
{
"chain_of_thought": "Step 1 - Profile Analysis: User specializes in AI safety research — RLHF, Constitutional AI, reward hacking, instrumental convergence. Step 2 - Response Analysis: Covers generic production risks at a business level; no mention of alignment-specific failure modes. Step 3 - Alignment Assessment: Topically related but misses the researcher-level depth the user expects (e.g., deceptive alignment, distributional shift). Step 4 - Score Decision: Partial match.",
"score": 1,
"reasoning": "The response addresses surface-level risks but lacks the technical alignment-research depth the user's profile requires."
}

User:
Input Data:
<user_profile>
{user_profile}
</user_profile>
<question>
{question}
</question>
<response_to_evaluate>
{response}
</response_to_evaluate>
Follow the Chain-of-Thought steps above and return your evaluation as a single JSON object with fields: "chain_of_thought", "score" (integer 0/1/2), and "reasoning".

Figure A6: CoT Evaluation Prompt.

System:
Role: You are a fair and insightful judge with exceptional reasoning and analytical abilities. Your task is to evaluate whether the response to the user's question aligns with the user's factor profile.

Scoring Rubric (3-point Likert scale):

- 0 (Mismatch): The response actively violates the user's preference or completely ignores a high-priority requirement defined in the profile.
- 1 (Partial Match): The response addresses the requirement but lacks depth, or only partially aligns with the user's preference.
- 2 (Full Match): The response perfectly adapts to the user's constraints and preferences described in the profile.

User:
Input Data:
<user_factor_profile>
{factors_profile}
</user_factor_profile>

<question>
{question}
</question>

<response_to_evaluate>
{response}
</response_to_evaluate>

The response should be in JSON format:

```
{
   "factor": "",
   "score": 0,
   "reasoning": "Brief explanation..."
},
{
   "factor": "",
   "score": 0,
   "reasoning": "Brief explanation..."
},
...
```

Requirements:
1. Each factor in the response corresponds to a factor in the user_factor_profile. Please ensure that every factor has a result.
2. Analyze the match between the <user_factor_profile> and <response_to_evaluate> for EACH factor.
3. Assign a score of 0, 1, or 2. **Important: The 'score' field must be a raw INTEGER type (int), do not use strings.**
4. Provide a brief reasoning for your score.
5. Output the result in strict JSON format.

Figure A7: Random-Factor Evaluation Prompt.

Your response must incorporate personalization by addressing the following six dimensions:

1. **Cognitive Trust:** Does the response align with the user's threshold for trust and credibility within this specific domain?
2. **Situational Anchoring:** Is the response precisely calibrated to the user's immediate context and specific problem?
3. **Schema Consistency:** Does the response integrate coherently with the user's prior knowledge and existing mental models?
4. **Cognitive Load Management:** Is the complexity of the response tailored to match the user's cognitive capacity?
5. **Metacognitive Scaffolding:** Does the response provide structural support that fosters the user's critical thinking skills?
6. **Affective and Motivational Resonance:** Does the response resonate with the user's current emotional state and motivational orientation?

Figure A8: Factors Prompt.

System:
Role: You are an intelligent assistant skilled in teaching. Your task is to generate personalized answers to user questions.

User:
User's Question:
{question}

The response should be in JSON format:
{
"answer": "",
"reasoning": "Brief explanation..."
}
Requirements:
1. The answer needs to be accurate.
2. Reasoning is a brief explanation of how you arrived at the answer above.
3. Ensure the output adheres to the specified format.

Figure A9: No-Personalization Prompt.

System:
Role: You are an intelligent assistant skilled in teaching. Your task is to generate responses tailored to the user's individual understanding, based on their question history.

User:
User's Historical Questions:
{Historical_questions}
User's Current Question:
{Current_question}

The response should be in JSON format:
{
"answer": "",
"reasoning": "Brief explanation..."
}
Requirements:
1. The answer needs to be accurate.
2. Reasoning is a brief explanation of how you arrived at the answer above.
3. Ensure the output adheres to the specified format.

Figure A10: Time-Personalization Prompt.

System:
Role: You are an intelligent assistant skilled in teaching. Your task is to generate responses tailored to the user's individual understanding, based on their question history.

User:
User's Historical Questions:
{Historical_questions}
User's Current Question:
{Current_question}

The response should be in JSON format:
{
"answer": "",
"reasoning": "Brief explanation..."
}
Requirements:
1. The answer needs to be accurate.
2. Reasoning is a brief explanation of how you arrived at the answer above.
3. Ensure the output adheres to the specified format.

Figure A11: RAG-Personalization Prompt.

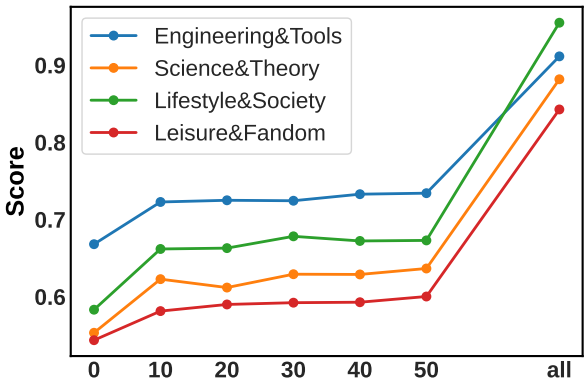


Figure A12: Impact of User History Context Length (K) on Model Performance

System:
Role: You are an intelligent assistant.

User:
question:
{question}
Please tell me which domain of the question is related to.
Requirements:
1. Please provide your decision in JSON format, following this structure:
{
"domain": "A summarization of which domain this question is related to" (if you are unable to summarize it, please set this value to "None"),
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure 'domain' is a single word.
3. The "reasoning" has no word limits.
4. Do not provide any other text outside the JSON string.

Figure A13: Domain Extract Prompt.

System:
You are an expert in {domain}. Please tell me the profile the user who asked the question in {domain}.

User:
question:{question}
Requirements:
1. Please provide your summary in JSON format, following this structure:
{
"profile": "A Summary of the user's profile in this domain"(starting with "This user"),
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure that the "profile" is accurate.
3. Do not provide any other text outside the JSON string.

Figure A14: Domain Profile Extract Prompt.

System:
You are an expert in {domain}. Please generate a new user profile based on the user's historical profile and the current profile.

User:
history user profile:{history}
current user profile:{current}

Requirements:
1. Please provide your output in JSON format, following this structure:
{
"profile": "A Summary of the user's profile in this domain"(starting with "This user"),
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure that the "profile" is accurate.
3. Do not provide any other text outside the JSON string.

Figure A15: Domain Profile Synthesize Prompt.

System:
You are an intelligent assistant. Please summarize the user profile based on the information about the user in various domains.

User:
user global profile: {global_profile}
user domain profile: {domain_profile}

Requirements:
1. Please provide your output in JSON format, following this structure:
{
"profile": "A Summary of user profile" (starting with "This user"),
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure that the "profile" is comprehensive and accurate.
3. Do not provide any other text outside the JSON string.

Figure A16: Global Profile Generate Prompt.

System:
Role: You are an intelligent assistant skilled in teaching. Your task is to generate personalized answers to user questions.

User:
User's Profile:
{user_profile}
User's Current Question:
{Current_question}

The response should be in JSON format:
```
{
   "answer": "",
   "reasoning": "Brief explanation..."
}
```
Requirements:
1. The answer needs to be accurate.
2. Reasoning is a brief explanation of how you arrived at the answer above.
3. Ensure the output adheres to the specified format.

Figure A17: Answer Generation Prompt.

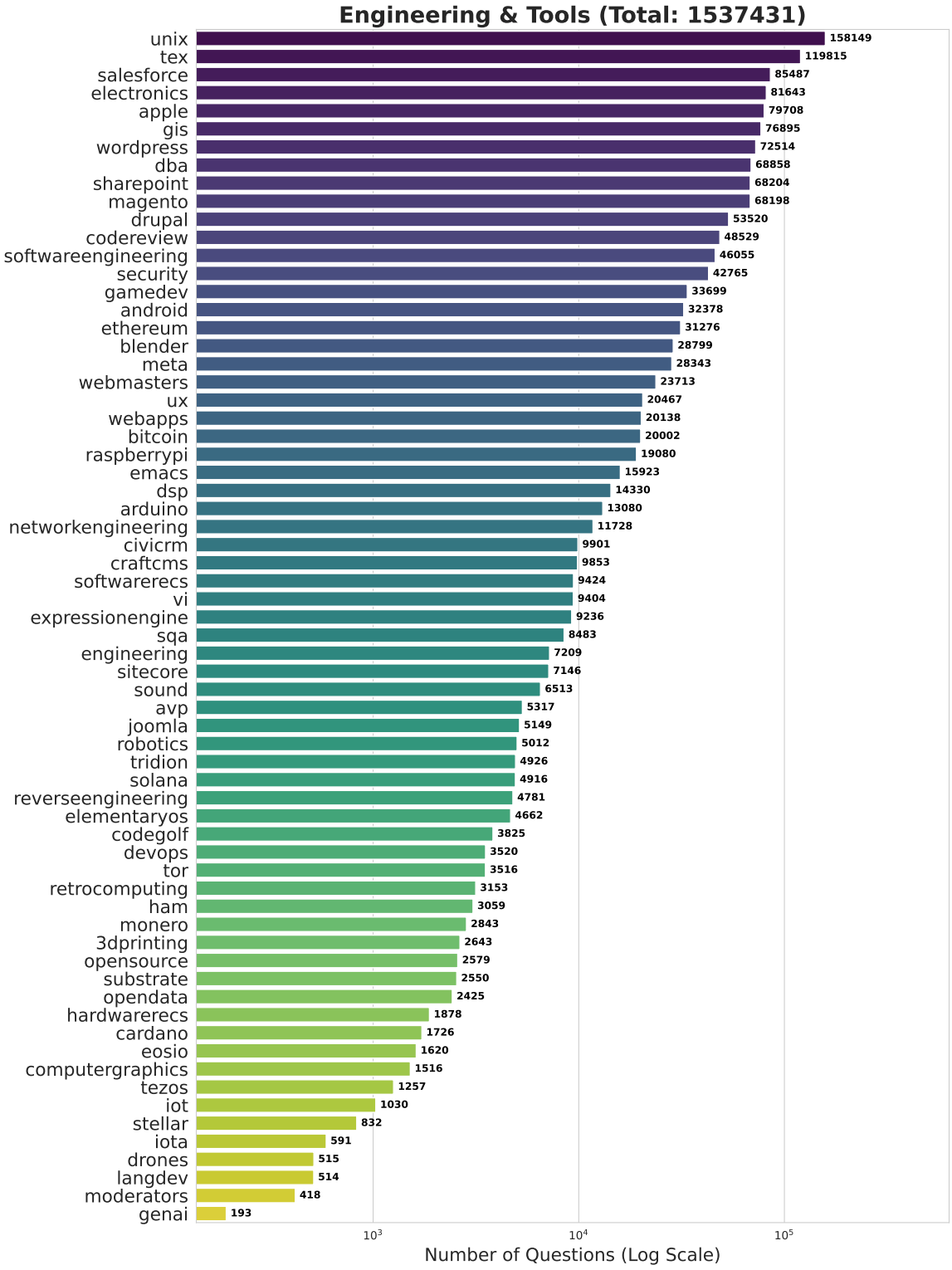


Figure A18: The distribution of questions across various domains in the Engineering&Tools.

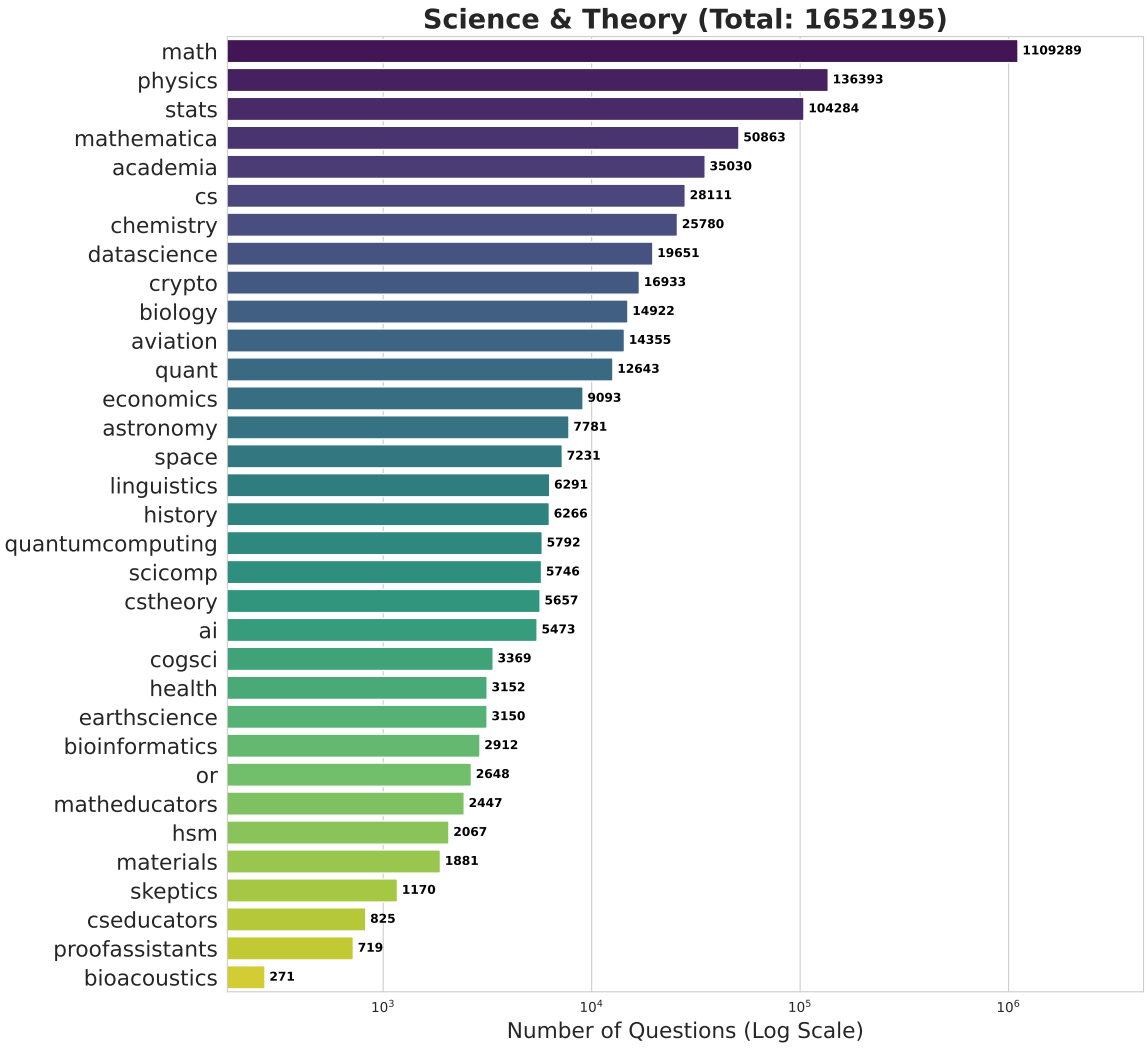


Figure A19: The distribution of questions across various domains in the Science&Theory.

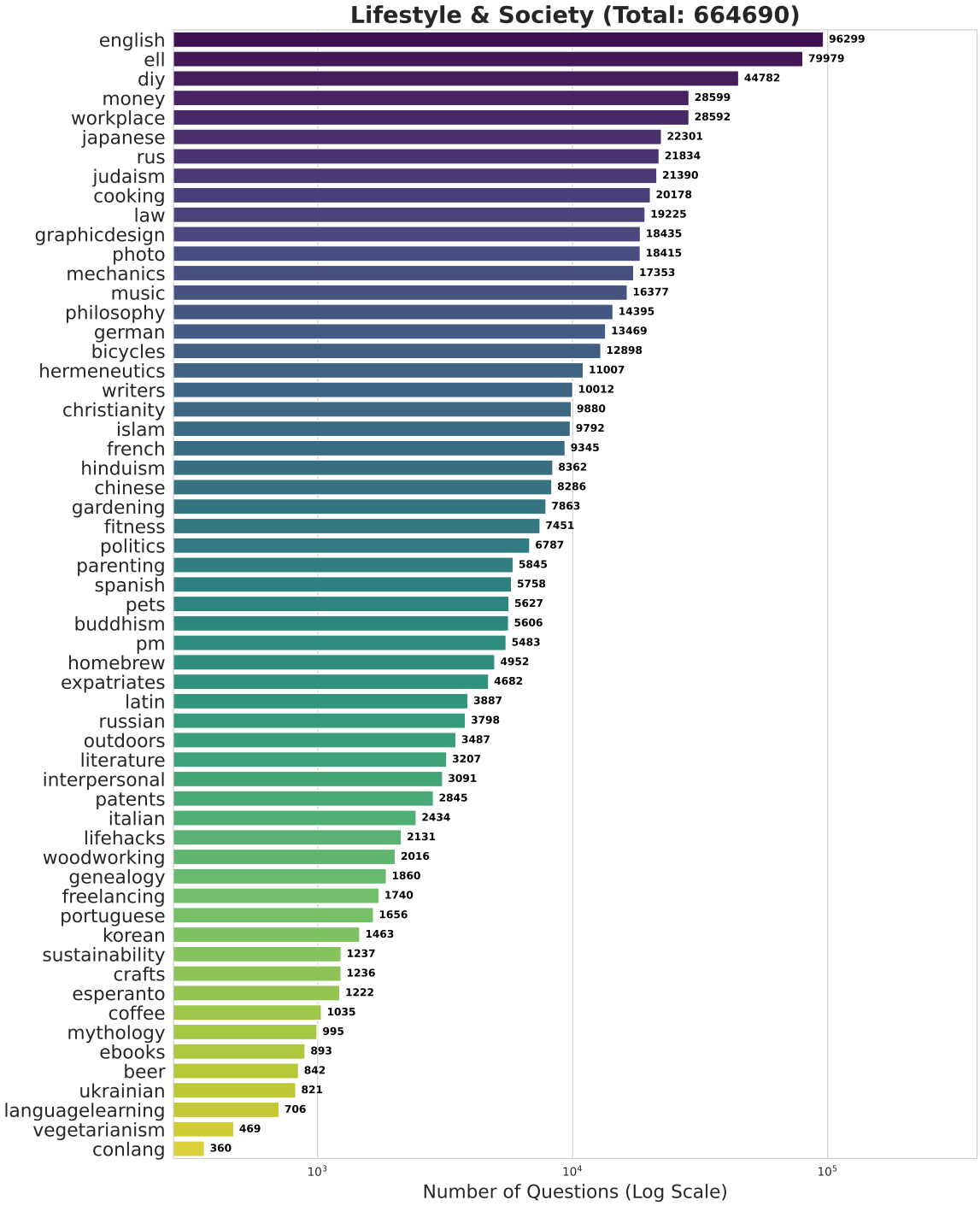


Figure A20: The distribution of questions across various domains in the Lifestyle&Society.

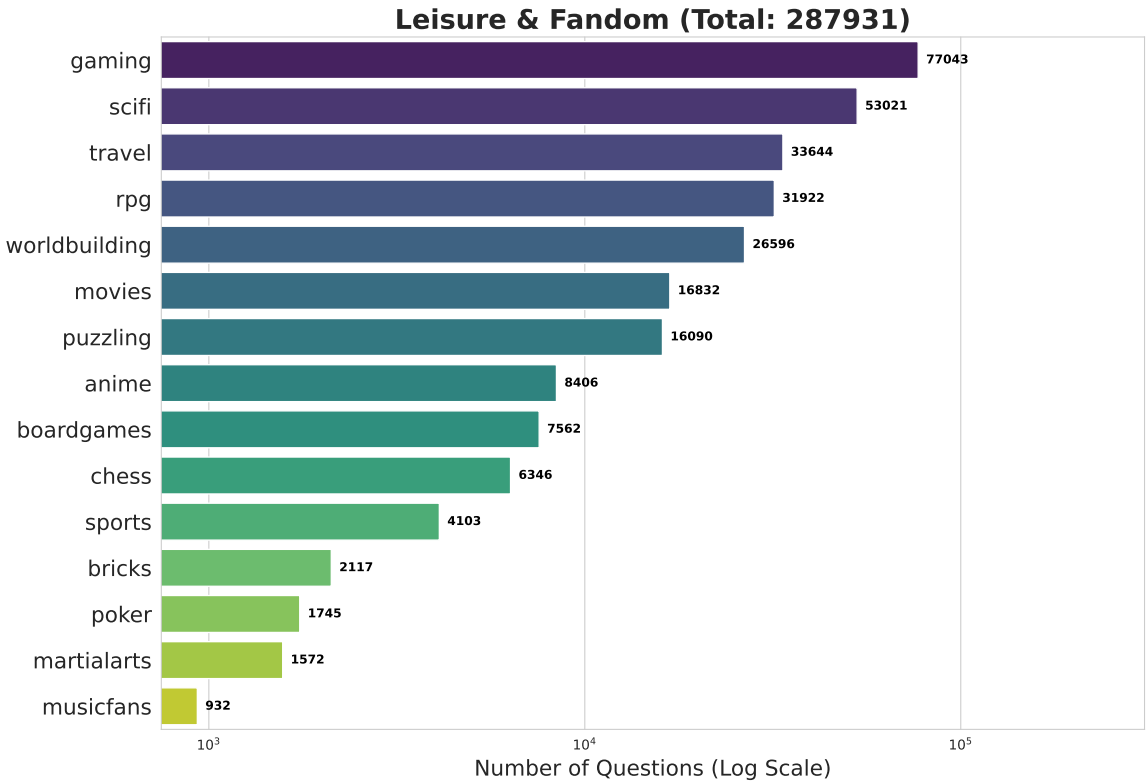


Figure A21: The distribution of questions across various domains in the Leisure&Fandom.