LLM-as-an-Explainer: Evaluating and Aligning LLM-generated Explanations for Scientific Concepts

Anonymous ACL submission

Abstract

People are increasingly using Large Language Models (LLMs) to explain unfamiliar scientific concepts. However, it is unclear whether LLM-generated explanations are accurate, clear, and useful. In this paper, we investigate *LLM-as-an-explainer* by (1) evaluating the quality of LLM-generated concept explanations, and (2) aligning open-source LLMs to produce *high-quality* concept explanations. In particular, we collect a large-scale dataset of 31,160 explanations generated by ten LLMs covering concepts from six disciplines, including Social Science, Biomedical Science, Mental Health, Computer Science, Law and Policy, and Finance. Next, we design a principle-guided evaluation framework that systematically assesses the quality of LLM-generated explanations. Our human validation shows substantial agreement between the proposed evaluation framework and human results. Finally, we propose ExpDPO to align lightweight LLMs by learning from multi-level *good* and *bad* paired concept explanations. Experiments show that the aligned LLMs can outperform their larger variants on this task.

1 Introduction

Recent Large Language Models (LLMs) have shown great proficiency in generating human-level responses. With the vast amount of human knowledge gleaned from training data, they are becoming primary sources for people seeking explanations of various scientific concepts (Xiong et al., 2024; Wu et al., 2024). As shown in Figure 1, *asking LLMs* appears to be a convenient and efficient fashion compared to traditional online searching. However, the lack of targeted evaluation makes the quality of LLM-generated concept explanations unclear. To bridge the gap, we pioneer the exploration of *LLM-as-an-explainer*.

In fact, generating a "high-quality" explanation for an abstract scientific concept is a challenging task. *Bad* explanations can mislead people, potentially causing misunderstanding and harm (Lakkaraju and Bastani, 2020). In contrast, a *good* explanation requires more than just factual accuracy; it should be easy to follow with concise language and real-world examples (Miller, 2019). To systematically investigate existing LLMs' capabilities in explaining concepts, we follow a comprehensive research pipeline (see Figure 2), beginning with the creation of a **large-scale dataset**, collecting 31,160 explanations from ten LLMs for 3,116 scientific concepts. Those concepts are chosen from Wikipedia, online dictionary, and related literature in six disciplines, including Social Science, Biomedical Science, Mental Health, Computer Science, Law and Policy, and Finance (Diener, 2000; Topp et al., 2015; TOV, 2018). We then propose a novel **evaluation framework** that adapts the principle-guided evaluation paradigm, using distinct Large Reasoning Models (LRMs) guided with fine-grained principles, such as "Utility" and "Accessibility", to thoroughly assess explanation quality. Finally, we investigate **pathways for improvement** by aligning lightweight open-source models using Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) (Rafailov et al., 2023). Moreover, we propose ExpDPO, a DPO-based method that employs a tailored reweighting mechanism to align high-performing concept explainer models by balancing *extreme* preference pairs (to preserve strong signals) and *hard-negative* pairs (to refine subtle differences).

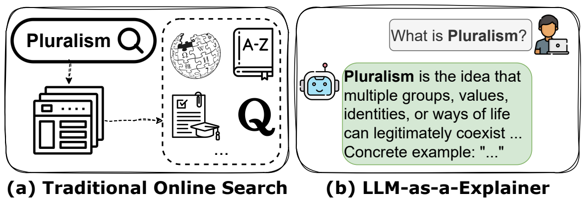

Figure 1: Illustration of concept explanation using (a) traditional online search and (b) LLM-as-an-Explainer.

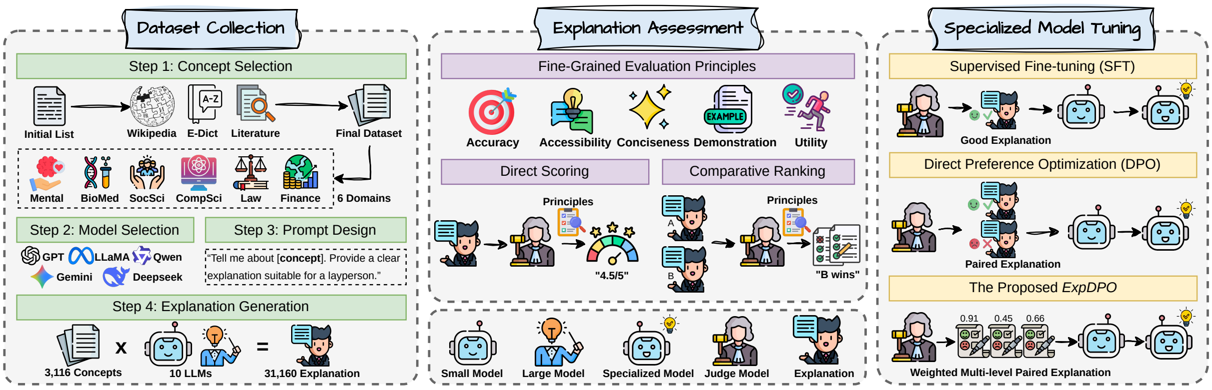

Figure 2: Overview of the research pipeline. From left to right, we first collect a large-scale LLM-generated concept explanation dataset. We then design fine-grained evaluation principles to assess the quality of collected explanations. Finally, we align specialized lightweight models to generate high-quality concept explanations.

Our empirical results first validate that the principle-guided evaluation framework provides a reliable assessment that aligns with human evaluators. Our analysis of the ten LLMs reveals performance disparities for explaining concepts across different domains. Most open-source LLMs show a relatively high likelihood of generating factually inaccurate explanations. Moreover, among the best-performing models, such as o4-mini (OpenAI, 2025) and Gemini-2.5-flash (Comanici et al., 2025), exhibit shared weaknesses in providing practical advice. To address these issues, we annotate sets of *good* and *bad* explanations and propose ExpDPO for improved model alignment of better scientific concept explanation. We empirically demonstrate that ExpDPO substantially improves the quality of the concept explanation in different settings. As illustrated in Figure 3 (examples of *sociocentrism*), compared to the larger base variants (Qwen-3-8B and 14B), the lightweight aligned LLM (Qwen-3-4B-ExpDPO) can generate a high-quality explanation that is close to the expected patterns of a *good* scientific concept explanation. Our main contributions are as follows:

- **Novel Datasets:** We develop a large-scale LLM-generated concept explanation dataset. It consists of 31,160 explanations generated by ten LLMs from six domains.
- **Principle-guided Evaluation:** We design an evaluation framework with fine-grained principles. We evaluate each concept explanation using both direct scoring and comparative ranking.
- **Empirical Experiments:** We propose ExpDPO to train lightweight explanation models. We conduct comprehensive experiments to show the concept explanation quality improvements.
- **Practical Implications:** We provide in-depth analyses on the effects of principles, domains, and models. We discuss practical implications on how to leverage LLM-as-an-explainer for various downstream tasks.

2 Related Work

2.1 LLM-generated Explanations

LLM-as-an-Explainer has been increasingly used to generate rationales and examples that make technical or abstract content more accessible. In the domain of Explainable Artificial Intelligence (XAI), recent work explored using LLMs to transfer ML explanation into human-readable narratives that are more interpretable (Zytek et al., 2024). In the clinical domain, a large-scale evaluation (Guo et al., 2025b) found that LLM-generated plain language summaries of clinical notes can be rated highly on perceived readability, but still underperform human-written text in enhancing layperson comprehension. Moreover, Yu et al. (2025) examined LLMs' understanding of physical concepts at different levels. Despite these advances, existing studies lack comprehensive, cross-domain benchmarks for systematically evaluating concept explanation quality and robustness. Our work addresses this gap by providing a large-scale benchmark dataset for concept explanations across diverse domains and providing evaluations of ten popular LLMs.

2.2 Evaluation of LLM-generated Content

Traditional Evaluation Metric. Traditional metrics like BLEU (Papineni et al., 2002) and

Concept Explanations of "Sociocentrism"

Qwen-3-4B. *Sociocentrism keeps societies running by encouraging people to follow shared rules. For example, someone might choose a career path because their community values it, like a girl wanting to be a doctor to meet her family's expectations, even if her passion is different.*

Qwen-3-8B. *Understanding sociocentrism helps us recognize biases that can divide people. For example, someone might think their country's traditions are superior, ignoring the value of other cultures, which can create misunderstandings or conflict. Acknowledging this bias fosters empathy and open-mindedness.*

Qwen-3-14B. *Sociocentrism can lead to prejudice because it makes people see their group as the 'norm' and others as 'different' or 'inferior.' For example, someone from a small town might think city life is chaotic, while a city dweller might view rural traditions as outdated — both sides failing to see the value in each other's perspectives.*

Qwen-3-4B-ExpDPO. *Sociocentrism is a kind of group-centered bias.* **[concise definition]** *We all grow up learning what our family, community, or society treats as normal and correct, so it's easy to use those rules as the standard for everyone else.* **[simple language]** *When that happens, we may automatically favor our own group's interests or assume our group's customs are "just how things should be," while other groups' ways feel odd or wrong.* **[demonstration]** *This doesn't necessarily come from hostility, it often comes from familiarity, but it can create real problems, like miscommunication, stereotyping, and unfair treatment.* **[implication]** *A useful habit is to notice when you're using your group as the default, ask what a behavior means in someone else's context, and judge actions with more than one social lens.* **[practical advice]**

Figure 3: Comparison of Qwen-3-4B, 8B, 14B, and 4B-ExpDPO explanations of "sociocentrism". Highlighted parts of the explanation generated by Qwen-3-4B-ExpDPO indicate the components that satisfy the evaluation principles.

ROUGE (Lin, 2004) rely heavily on exact matching to evaluate models' generation quality. Subsequent methods, such as BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), improve upon this by using contextual embeddings, but remain incapable of capturing nuanced features (Post, 2018).

LLM-based Evaluation. The advanced capabilities of LLMs have inspired a paradigm shift towards dynamic reference-free assessment (Wang et al., 2023). LLMs-as-evaluators, as one of the leading automated evaluation paradigms, has been widely adopted due to its ability to conduct nuanced evaluations like humans (Zheng et al., 2023; Li et al., 2024). It has been used in domains like academic writing (Liu and Shah, 2023), code generation (McAleese et al., 2024), and computational social science (Jiang et al., 2024), to evaluate the quality of open-ended generation. However, recent studies have revealed various biases and vulnerabilities of the LLMs-as-evaluators, raising concerns in this technique (Li et al., 2025).

Principle-guided Evaluation. To address these limitations, researchers proposed principle-guided LLM-based evaluation (Li et al., 2024), where a set of comprehensive and well-designed rules or rubrics is given to the LLM evaluator for improving the assessment's fairness and reliability. Following studies further improve it by providing domain (Ye et al., 2023) or sample-level principles (Kim et al., 2025; Gunjal et al., 2025; Viswanathan et al., 2025), instructing LLM evaluators with more fine-grained guidelines. Building on this line of work, we introduce evaluation principles specific to concept explanation: tailored principles that align the evaluator's perspective with the needs of general users.

3 Research Pipeline

3.1 Collecting Concept Explanation Dataset

To systematically evaluate the quality of LLM-generated concept explanations, we conduct a rigorous data collection procedure comprising the following steps:

Concept Selection. We start by compiling a comprehensive list of concepts across six domains: Social Science, Biomedical Science, Mental Health, Computer Science, Law and Policy, and Finance. Initial concepts are identified based on their relevance, popularity, and coverage in related literature (Diener, 2000; Topp et al., 2015; TOV, 2018). We further expand this list through cross-referencing synonyms and related terms from Wikipedia and the Oxford English Dictionary. The final dataset consists of **732** Social Science, **1,011** Biomedical Science, **451** Mental Health, **407** Computer Science, **299** Law and Policy, and **216** Finance-related concepts.

Model Selection. We select ten LLMs for generating concept explanations. These include four larger API-based proprietary models known for their advanced capabilities:

- GPT-4.1-mini (Achiam et al., 2023).
- OpenAI-o4-mini (OpenAI, 2025).
- Gemini-2.5-flash (Comanici et al., 2025).
- Deepseek-V3 (Liu et al., 2024).

We also select six smaller open-source LLMs:

- Qwen-3 (1.7B, 4B, 8B, and 14B) (Yang et al., 2025).
- LLaMA-3.2-instruct (1B and 3B) (Grattafiori et al., 2024).

This combination provides comprehensive coverage across different model sizes, architectures, and training paradigms.

User Prompt Selection. To ensure consistency and emulate realistic user-LLM interactions, we select a standardized prompt template:

> *"Tell me about [concept]. Provide a clear explanation suitable for a layperson."*

We test variants of concept explanation prompts. Given the similar responses provided by them, we use the above one to generate all concept explanations. More details can be found in Appendix C.

Concept Explanation Generation. Applying the standard prompt template, we query each of the 3,116 concepts using all selected LLMs, resulting in a total of 31,160 concept explanations. To minimize variability and randomness in model outputs, all generations are conducted by setting LLMs' `temperature` $= 0$ if available.

3.2 Assessing Concept Explanation Quality

We employ a principle-guided LLM-based evaluation framework to assess concept explanation quality, leveraging two powerful Large Reasoning Models (LRMs) as judging LLMs, $J = \{J_1, J_2\}$, where J_1 is Gemini-2.5-Pro (Comanici et al., 2025) and J_2 is DeepSeek-R1 (Guo et al., 2025a). They will assess the quality using *Direct Scoring* and *Comparative Ranking* based on five predefined evaluation principles. Our prompt templates for evaluation are included in Appendix D

Fine-Grained Evaluation Principles. To enhance consistency and interpretability of the evaluation process, inspired by previous work (Ye et al., 2023), we carefully design five evaluation principles tailored for broad usability:

- Accuracy: Provide an accurate definition of the concept.
- Accessibility: Use of simple, everyday language.
- Conciseness: Brief and direct explanations without unnecessary verbosity.
- Demonstration: Use of relatable analogies, stories, or real-world examples.
- Utility: Provision of actionable and practical advice, or implication.

Direct Scoring. Judging LLMs assign a score to each explanation *per principle*. For a given concept explanation $E_{i,j}$ (for concept c_i and generation model M_j), each judging LLM J_l provides a score $S_l(E_{i,j}, v) \in [1, 5]$ for each principle v. The final score for an explanation on a given principle is the average score from J_1 and J_2:

$$S(E_{i,j}, v) = \frac{1}{2} \sum_{l=1}^{2} S_l(E_{i,j}, v). \quad (1)$$

To assess a model's performance on a specific principle, we average the scores $S(E_{i,j}, v)$ across all concepts C. The mean quality score for model M_j on principle v is:

$$Q(M_j, v) = \frac{1}{|C|} \sum_{i=1}^{|C|} S(E_{i,j}, v). \quad (2)$$

For *Accuracy*, we prompt LLMs with a concept definition from Wikipedia, Oxford English Dictionary, or related literature as a reference, and instruct them to score closeness relative to that definition.

Comparative Ranking. Two judging LLMs compare each explanation against a baseline reference explanation E_{ref} (i.e., Qwen-3-4B) *per principle*. For each principle v, the comparison yields an outcome $O(E_{i,j}, v) \in \{\text{win, loss}\}$. We involve a third judge model J_3 (i.e., GPT-5.2) only when there is a conflict between J_1 and J_2 (e.g., J_1 outputs *win* and J_2 outputs *loss*). The final result is decided by majority vote. The win rate for model M_j on principle v is calculated as:

$$W(M_j, v) = \frac{|\{E_{i,j} \mid O(E_{i,j}, v) = \text{win}\}|}{|C|} \tag{3}$$

For *Accuracy*, the LLM assigns an outcome based on which explanation (E_{ref} and $E_{i,j}$) is closer to the online definition. For example, if the baseline explanation E_{ref} is closer to Wikipedia's definition, the final outcome will be *loss*.

Human Evaluation To validate the reliability of the proposed evaluation framework, we conduct a human evaluation using Direct Scoring and Comparative Ranking. Specifically, we compare the LLMs' evaluations against human annotations on 20 randomly selected explanations for each discipline (i.e., 1,200 explanations in total). We recruit three annotators (all Ph.D. students) and develop an annotation tool with a simple UI. Annotators score each explanation using a 1–5 scale and provide pairwise preferences (win/loss) on the **same evaluation principles** with randomized presentation order and hidden model identities to reduce bias. We compute the inter-rater agreement using Cohen's kappa (Cohen, 1960). More details are presented in Appendix A.

3.3 Aligning Specialized Model

We align lightweight LLMs to generate high-quality concept explanations using Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO). We further propose **EXPDPO** tailored for our annotated preference data. These methods are applied separately to the same pre-trained base model (i.e., Qwen-3-4B) using the HuggingFace Transformers library[1]. We employ the ZeRO-3[2] from the DeepSpeed library for training. We denote the base model being trained as M_θ. All models are fine-tuned on two A100 GPUs. More details can be found in Appendix B.

3.3.1 Supervised Fine-Tuning.

SFT aims to adapt the base model M_θ to generate outputs that conform to the format and style of high-quality explanations.

[1] https://huggingface.co/Qwen/Qwen3-4B

[2] https://deepspeed.readthedocs.io/en/latest/zero3.html

SFT Data Preparation. We construct the SFT dataset D_{SFT} by applying a filtering process to each concept explanation to select *good* responses. For each concept, an explanation (E_i) will be included in D_{SFT} if it is ranked top-2 from the direct scoring (descending order of overall scores) described in Section 3.2.

SFT Objective. The base model M_θ is fine-tuned on the curated dataset D_{SFT}. For each concept, we use the standardized prompt template containing the concept c_i as the input prompt P_i and the corresponding explanation as the target output E_i. The SFT objective is to train M_θ by minimizing the negative log-likelihood loss $\mathcal{L}_{\text{SFT}}(\theta)$:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{\substack{(P_i,E_i)\\ \in D_{SFT}}} \sum_{t=1}^{|E_i|} \log P(E_{i,t}|P_i, E_{i,<t}; \theta) \tag{4}$$

where $E_{i,t}$ is the t-th token of the target explanation E_i.

3.3.2 Direct Preference Optimization.

DPO aligns the policy model using pairwise preferences (Rafailov et al., 2023), encouraging the model to assign higher likelihood to *good* explanations than to *bad* ones.

DPO Data Preparation. We create two sets of preference pairs: *extreme pairs* (D_{ext}) and *hard-negative pairs* (D_{hard}).

Extreme pairs. For each concept, we select extremely *good* and *bad* explanations:

- $E^g_{princple}$: the best explanation according to the direct scoring for $[principle]$.
- $E^b_{principle}$: the worst explanations according to the direct scoring for $[principle]$.

For example, a pair of explanations ($E^g_{accuracy}$, $E^b_{accuracy}$) for concept c_i contains the best and worst explanations based on *accuracy* scores.

Hard-negative pairs. To encourage fine-grained improvements, we select hard negative samples that are difficult to distinguish from the good one:

- $E^g_{principle}$: the best explanation according to the direct scoring for $[principle]$.
- $E^b_{principle}$: the near-top explanation (rank 3 or 4) according to the direct scoring for $[principle]$.

The final dataset is $D_{\text{DPO}} = D_{\text{ext}} \cup D_{\text{hard}}$, containing multi-level preference data.

DPO Objective. DPO optimizes the policy model to increase the likelihood of explanation E^g over E^b, guided by a frozen reference model (the original base model). The standard DPO loss is:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(P,E^g,E^b)} \left[\log \sigma \left(r_\theta(P, E^g) - r_\theta(P, E^b)\right)\right], \tag{5}$$

where $r_\theta(P, y) = \beta \log \frac{\pi_\theta(y|P)}{\pi_{ref}(y|P)}$ is the implicit reward function, π_θ is the policy model, π_{ref} is the reference model, $\sigma(\cdot)$ is the sigmoid function, and β controls the deviation from the reference model.

The proposed ExpDPO. As discussed before, our preference data contains two types of pairs. We propose ExpDPO to jointly optimize over both types using a margin-weighted mixture objective:

$$\mathcal{L}_{\text{ExpDPO}}(\theta) = -\sum_{t \in \{\text{ext,hard}\}} \lambda_t \, \mathbb{E}_{(P,E^g,E^b)\sim D_t} \left[\alpha_t(\Delta) \cdot \log \sigma\left(r_\theta(P, E^g) - r_\theta(P, E^b)\right)\right], \tag{6}$$

where D_{ext} and D_{hard} denote the sets of extreme and hard-negative pairs, respectively, and $\lambda_{\text{ext}} + \lambda_{\text{hard}} = 1$ controls their relative contribution. For each pair associated with principle v, we define a score margin from direct scoring as:

$$\Delta = S(E_g, v) - S(E_b, v), \tag{7}$$

where $S(\cdot, v) \in [1, 5]$ is the direct score for principle v, and thus $\Delta \in [0, 4]$. We set $\alpha_t(\Delta)$ to reflect the different roles of the two pair types: (i) extreme pairs provide high-confidence signals and are emphasized when the margin is large, while (ii) hard-negative pairs provide fine-grained supervision and are emphasized when the margin is small. Concretely, we use $\alpha_{\text{ext}}(\Delta) = \left(\frac{\Delta}{4}\right)^{\gamma_{\text{ext}}}$ and $\alpha_{\text{hard}}(\Delta) = \left(\frac{1}{\max(\Delta,\epsilon)}\right)^{\gamma_{\text{hard}}}$, where $\epsilon > 0$ avoids division-by-zero when $\Delta = 0$, and γ_{ext} and γ_{hard} control the sharpness of reweighting. We set $\epsilon = 0.5$ in all experiments.

4 Experiments and Results

In this section, we present our empirical results to answer the following research questions:

- **RQ1**: Does the principle-guided evaluation framework provide human-level evaluation?
- **RQ2**: How do the concept explanation capabilities among the evaluated LLMs?
- **RQ3**: To what extent can the proposed ExpDPO improves LLM-generated concept explanation?

4.1 Human Validation of the Principle-guided Evaluation Framework (RQ1)

Figure 4 presents the level of agreement between LLMs and human annotators. We observe that results from the principle-guided evaluation framework are very close to human evaluation in both evaluation strategies. Moreover, there is no significant inter-rater agreement discrepancy between the six disciplines. This validates that our evaluation framework can serve as a **reliable proxy** of human annotators for evaluating multidisciplinary scientific concept explanations.

4.2 Differences in Concept Explanation Capability (RQ2)

We conduct comprehensive analyses of the evaluated LLMs. Our results reveal significant disparities in LLMs' capabilities based on several factors.

✦ **Principle-wise Analysis.** LLMs present unified weakness in Utility and distinct strengths in other principles. As shown in Figure 5, while the larger LLMs consistently outperform smaller models across all evaluation principles, they exhibit a shared weakness in providing practical advice (*Utility*). At the same time, each of these models demonstrates particular strengths in specific principles. GPT-4.1-mini excels on *Accessibility*, o4-mini achieves the highest scores for factual *Accuracy*, DeepSeek-v3 is good at providing clear *Demonstration* and *Concise* explanations, and Gemini-2.5-flash can generate *Accurate* definitions as well. Although larger LLMs generally perform worse on *Utility* and *Depth*, Gemeni-flash-2.5 and DeepSeek-v3 demonstrate relatively better performances. Based on their overall performance (Figure 5), we list the *best* and *second best* models for each evaluation principle.

- Accuracy: o4-mini and Gemeni-flash-2.5
- Accessibility: o4-mini and DeepSeek-v3
- Conciseness: GPT-4.1-mini and o4-mini
- Demonstration: o4-mini and GPT-4.1-mini
- Utility: o4-mini and Gemeni-flash-2.5

✦ **Domain-wise Analysis.** Generating high-quality concept explanations for *social science, mental health, biomedical science,* and *law and policy* is **relatively more difficult** for all evaluated LLMs. This is reflected in the domain-specific radar plots in Figure 5, where performance becomes "contracted" in these domains. In contrast, computer

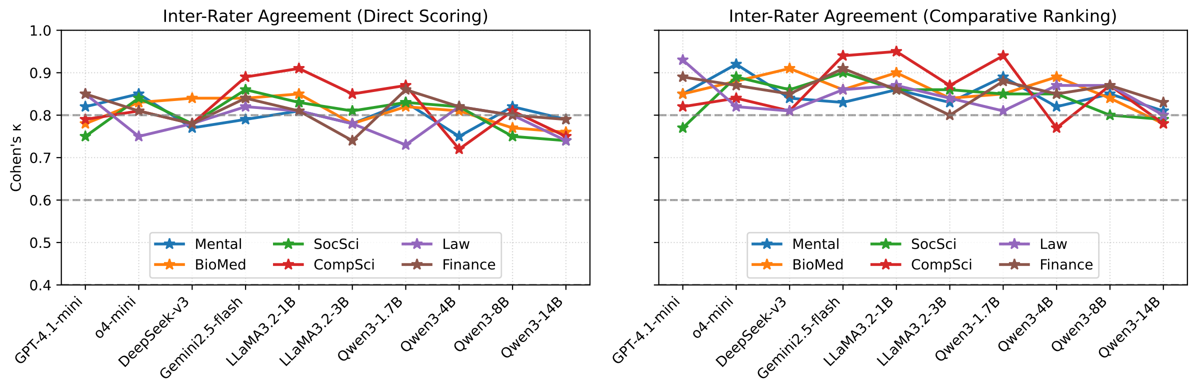

Figure 4: Cohen's kappa scores between LLM-based evaluation and human annotators. Dashed lines at 0.6 and 0.8 indicate substantial (0.61 to 0.80) and almost perfect (0.81 to 1) agreement, respectively.

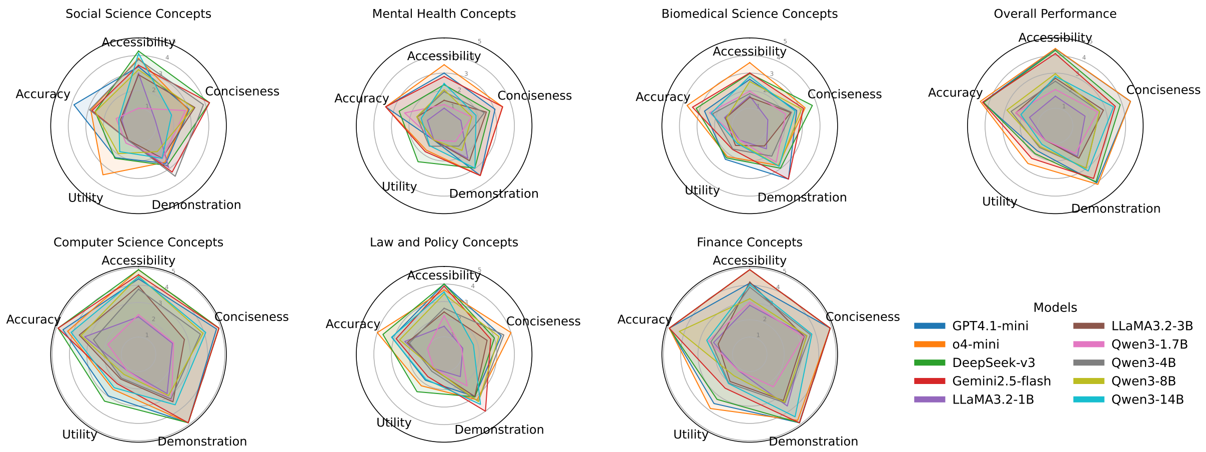

Figure 5: Direct Scoring comparisons of ten LLMs' concept explanations across six disciplines.

science and finance exhibit more balanced shapes across principles. Through our human verification of the generated explanation, we observe that (1) social science concepts are often theory-based and context-dependent, leading models to be hard to produce actionable advice (i.e., lower Utility); (2) mental health concept explanations require precise boundaries to avoid overgeneralization; (3) biomedical concepts often require terminology discipline, where models suffer from explaining them in concise languages; and (4) law and policy concepts contain jurisdictional nuances, where models may remain fluent and readable but miss key legal exceptions. These results suggest that improving concept explanation quality in social science, mental health, biomedical science, and law and policy likely requires *additional domain-specific data*.

✦ **Model-wise Analysis.** We observe that the four larger API-based LLMs (GPT-4.1-mini, o4-mini, DeepSeek-v3, and Gemini-2.5-flash) form substantially larger radar polygons than the smaller open-source models, indicating a **clear scale effect** as expected. In Table 1, within the open-source models, we see a monotonic quality improvement with model size. Notably, scaling shows relatively larger gains on Accuracy and Accessibility, suggesting that larger LLMs are better at generating accurate definitions with simple language. In other words, smaller models tend to generate shallow or inaccurate explanations. Moreover, some smaller models provide fewer concrete examples (e.g., Qwen3-1.7B), whereas others generate richer demonstrations at the cost of verbosity (e.g., LLaMA3.2-1B).

4.3 Improvements of Specialized Concept Explanation Models (RQ3)

We fine-tune the Qwen-3-4B base model using SFT, DPO, and ExpDPO. We compare their performance back on the same evaluation set. In particular, we begin with a pool of 1,200 concepts (200 per domain), split into 960 for training and 240

Model	Mental	BioMed	SocSci	CompSci	Law	Finance	Overall
Larger API-based Models							
GPT-4.1-mini	88.5	**92.3**	84.7	**98.5**	83.1	97.2	90.7
o4-mini	87.4	90.8	85.3	96.3	**85.8**	**98.4**	**90.7**
DeepSeek-v3	**89.1**	91.7	**85.9**	89.3	80.5	89.6	87.7
Gemini-2.5-flash	86.2	91.8	83.8	91.2	85.2	87.2	87.6
Smaller Open-source Models							
LLaMA-3.2-1B-Instruct	12.4	18.7	7.5	22.3	18.6	23.4	17.2
LLaMA-3.2-3B-Instruct	35.2	55.3	25.1	35.2	41.0	42.1	39.0
Qwen-3-1.7B	22.7	48.3	13.6	28.3	22.5	20.4	26.0
Qwen-3-8B	65.0	80.5	53.2	76.2	69.2	78.2	70.4
Qwen-3-14B	**78.4**	**88.7**	**65.9**	**87.7**	**75.3**	**80.1**	**79.4**

Table 1: Comparative Ranking comparisons of LLMs' concept explanation. All results indicate win rates (%) against the Qwen-3-4B baseline. **Bold** and underline values indicate the best and second-best results, respectively.

Model	In-Domain	OOD
Larger API-based Models		
GPT-4.1-mini	4.07	-
o4-mini	4.11	-
DeepSeek-v3	**4.14**	-
Gemini-2.5-flash	4.09	-
Smaller Open-source Models		
LLaMA-1B-Inst.	1.90	-
LLaMA-3B-Inst.	2.46	-
Qwen-3-1.7B	2.10	-
Qwen-3-8B	2.75	-
Qwen-3-14B	**2.88**	-
Qwen-3-4B	2.73	-
Lightweight Aligned Models		
Qwen-3-4B-SFT	3.22 (+17.9%)	2.90 (+6.2%)
Qwen-3-4B-DPO	3.29 (+20.5%)	3.01 (+10.3%)
Qwen-3-4B-ExpDPO	**3.46 (+26.7%)**	**3.18 (+16.5%)**

Table 2: Direct Scoring results (averaged scores of all principles and domains) are presented. Relative performance gains over Qwen-3-4B are shown in parentheses.

held out for evaluation. Using the annotated SFT and DPO data described in Section 3.3, we train the base model and test with in-domain and out-of-domain (OOD) settings. For the in-domain setting, we train and test aligned models using data from all six domains. For OOD, we train the model on five domains and test it on the remaining domain.

✦ **ExpDPO is the most effective approach.** As shown in Table 2, SFT and DPO can already achieve substantial gains over the Qwen-3-4B model. Qwen-3-4B-SFT increases the in-domain score by 0.49 points (+17.9%) and the OOD score by 0.17 points (+6.2%), completely outperforming the Qwen-3-8B and 14B. Qwen-3-4B-ExpDPO improves even further, adding 0.73 points (+26.7%) for in-domain and 0.45 points (+16.5%) for OOD. Figure 3 illustrates a case study of LLM-generated explanations of *"sociocentrism"* (more examples are included in Appendix E). We can observe that Qwen-3-4B-ExpDPO can generate relatively high-quality explanations compared to the base models.

✦ **Stronger generalization to unseen domains.** Table 2 shows that Qwen-3-4B-ExpDPO generalizes beyond the training domains compared to SFT and DPO. This finding is consistent with the role of weighted multi-level preference-based optimization. ExpDPO explicitly contrasts multi-level *good* and *bad* explanations and encourages the model to internalize domain-agnostic subtle signals, which are less sensitive to domain shifts.

5 Conclusion and Future Work

In this paper, we systematically evaluate the quality of LLM-generated concept explanations. We build a large-scale scientific concept explanation dataset, develop a principle-guided evaluation framework, and propose ExpDPO for aligning lightweight models for better concept explanations. Our findings reveal shared weaknesses and unique strengths of LLMs' concept explanation capability. Finally, we demonstrate that using ExpDPO and our preference data can substantially improve the quality of generated concept explanations. Looking forward, LLM-as-an-explainer can bring several benefits. For normal users, it enables efficient concept learning and understanding without extensive online searching. For scientific research, high-quality explanations can serve as (1) a representation of additional context and knowledge (e.g., knowledge graph construction); (2) a synthetic training signal for data augmentation; and (3) a semantic supervision for tasks such as reasoning and retrieval.

Limitations

Our work has several limitations. First, our benchmark covers concept explanations from six scientific disciplines, but it does not cover other disciplines, cultures, or languages, and may underrepresent emerging concepts. Second, the actual outcomes of learning from LLM-generated explanations and online searching are not clear. Controlled human studies and field experiments are needed to assess real-world impacts. Third, we only test the effectiveness of SFT, DPO, and the proposed EXPDPO for this task. Other alignment methods, such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024), remain unclear.

Ethical Considerations

Our human validation is conducted by three annotators who are Ph.D. students and coauthors of this paper (Appendix A). No sensitive personal data is collected. We acknowledge that LLM-generated concept explanations may be biased and harmful. We provide content warnings and opt-outs to the annotators throughout the whole study. We obtained Institutional Review Board (IRB) approval from our university to conduct an annotation task with prolific annotators. To prevent unnecessary computational cost and promote reproducibility, we use lightweight open-source models for post-training alignment. We apply efficient fine-tuning techniques to reduce the GPU overload. We will release the full dataset, codes, and models to the public to facilitate future research.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ed Diener. 2000. Subjective well-being: The science of happiness and a proposal for a national index. *American psychologist*, 55(1):34.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yue Guo, Jae Ho Sohn, Gondy Leroy, and Trevor Cohen. 2025b. Are llm-generated plain language summaries truly understandable? a large-scale crowdsourced evaluation. *arXiv preprint arXiv:2505.10409*.

Bohan Jiang, Dawei Li, Zhen Tan, Xinyi Zhou, Ashwin Rao, Kristina Lerman, H Russell Bernard, and Huan Liu. 2024. Assessing the impact of conspiracy theories using large language models. *arXiv preprint arXiv:2412.07019*.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, and 1 others. 2025. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5877–5919.

Himabindu Lakkaraju and Osbert Bastani. 2020. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv: 2411.16594*.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others.

2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

OpenAI. 2025. Openai o3 and o4-mini system card. Technical report, OpenAI.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Christian Winther Topp, Søren Dinesen Østergaard, Susan Søndergaard, and Per Bech. 2015. The who-5 well-being index: a systematic review of the literature. *Psychotherapy and psychosomatics*, 84(3):167–176.

William TOV. 2018. Well-being concepts and components. *Handbook of subjective well-being*, pages 1–15.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. 2025. Checklists are better than reward models for aligning language models. *arXiv preprint arXiv:2507.18624*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11.

Siyi Wu, Julie YA Cachia, Feixue Han, Bingsheng Yao, Tianyi Xie, Xuan Zhao, and Dakuo Wang. 2024. " i like sunnie more than i expected!": Exploring user expectation and perception of an anthropomorphic llm-based conversational agent for well-being support. *arXiv preprint arXiv:2405.13803*.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung, and Jie Zhou. 2025. The stochastic parrot on llm's shoulder: A summative assessment of physical concept understanding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11416–11431.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. 2024. Llms for xai: Future directions for explaining explanations. *arXiv preprint arXiv:2405.06064*.

Appendix

A Human Validation

We validate the results of the proposed principle-guided LLM-based evaluation framework with human annotators.

A.1 Annotators

We use three annotators, all of whom are Ph.D. students in Computer Science and coauthors of this paper. The study involves only reading LLM-generated text and assigning ratings; we do not collect sensitive or personally identifying information. Since the annotators are authors, no monetary compensation is involved.

A.2 Annotation Tool

As shown in Figure 6 and 7 (using *Acceptance and Commitment Therapy* as an example), we developed an annotation tool with a simple UI for both evaluation strategies:

- **Direct scoring:** given a concept explanation, annotators assign ordinal scores (1–5) for each principle.
- **Comparative ranking:** given a concept and two corresponding explanations (from model A and B), annotators select `A better` or `B better` for each principle.

To reduce presentation bias, the tool hides model identities and randomizes the left/right order of A/B in pairwise comparisons.

A.3 Sampling Details

We validate on a subset of the benchmark to cover all evaluated domains and models.

Direct scoring. For each domain, we uniformly sample $n = 20$ concepts and evaluate the corresponding explanations. In total, $N_{\text{ds}} = 20 \times |\mathcal{M}| \times |\mathcal{D}|$ items.[3] Each explanation is annotated by three annotators.

Comparative ranking. For each domain, we sample $n = 20$ concepts and construct an explanation pair for each concept. The reference explanation from Qwen-3-4B will be compared against the other nine explanations. Therefore, we have $N_{\text{cr}} = 20 \times |\mathcal{D}| \times 9$ comparison pairs. Each pair is annotated by three annotators.

[3] In our experiments, $|\mathcal{M}| = 10$ and $|\mathcal{D}| = 6$, giving $N_{\text{ds}} = 1200$.

Score	Anchor description (applies to each principle)
1	Very Poor - Fails to meet the principle.
2	Poor - Significant shortcomings in meeting the principle.
3	Adequate - Meets basic requirements but could be improved.
4	Good - Mostly meets the principle with minor shortcomings.
5	Excellent - Fully meets the principle

Table 3: Rubric anchors for human and LLM annotators.

A.4 Evaluation Rubric

Annotators follow the same principle definitions as the LLM evaluator and use a shared rubric with anchors for 1–5 scores (Table 3). Before formal annotation, we conduct a short calibration round on a small set of examples to align interpretations of each principle, then finalize the rubric used for the main validation. Annotators are allowed to use any resources (except LLMs) to check the ACCURACY of LLM-generated concept explanations.

A.5 Human Results

Direct scoring. For each evaluated concept explanation, we averaged scores from all three annotators as the final score.

Comparative ranking. For each evaluated concept explanation, we take the result of the majority vote from three annotators as the final result.

A.5.1 Calculation of Cohen's κ.

Since both human and LLM evaluations produce real-valued scores on a 1–5 scale, we discretize them by rounding to the nearest integer before computing κ. Formally, for any score $s \in [1, 5]$, we compute $\tilde{s} = \text{round}(s) = \lfloor s + 0.5 \rfloor$. For example, $4.1 \to 4$ and $4.9 \to 5$. We apply the same rounding rule to both the human score and the LLM score, and then compute Cohen's κ on the final discrete labels $\tilde{s} \in \{1, 2, 3, 4, 5\}$.

B Implementation Details

SFT implementation:

- Batching: per-GPU batch size = 4, gradient accumulation = 2 (effective batch size 8).
- Optimizer: AdamW, initial Learning Rate = 1×10^{-5}, cosine decay scheduler with 10% warmup.
- Training: 3 epochs, BF16 enabled, logging every 10 steps, checkpoint every 5k steps.

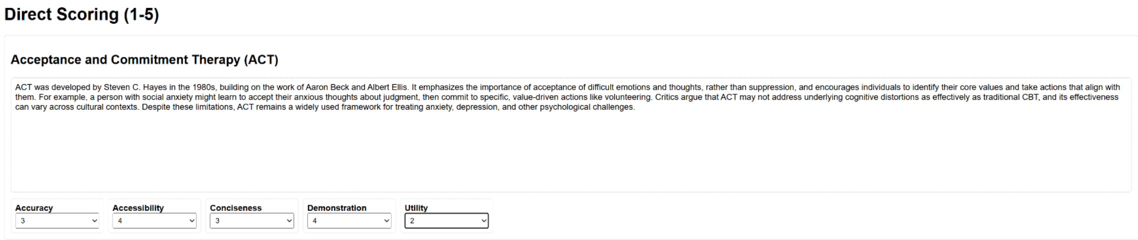

Figure 6: UI for Direct Scoring.

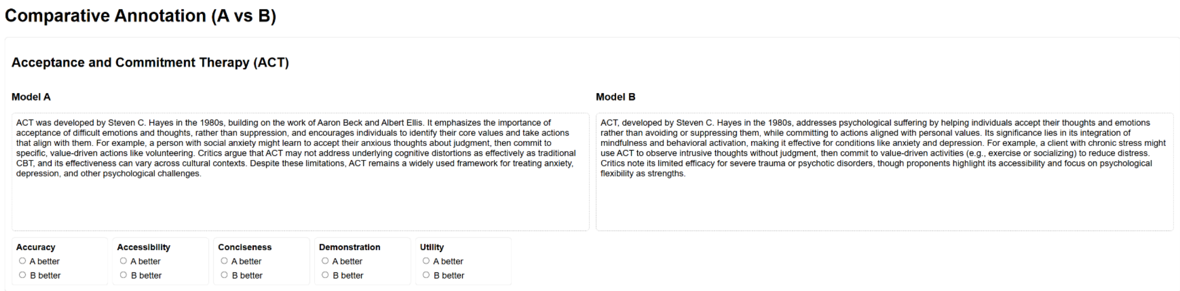

Figure 7: UI for Comparative Ranking.

DPO and ExpDPO implementation:

- Batching: per-GPU batch size = 1, gradient accumulation = 4 (effective batch size 4).
- Optimizer: AdamW, initial Learning Rate = 5×10^{-6}, cosine decay with 10% warmup.
- Training: 3 epochs, BF16 enabled, logging every 10 steps, checkpoint every 5k steps.

C Prompt Templates of Explanation Generation

This section details the alternative user prompt templates we tested in addition to the standardized prompt used in the main experiments. All prompt templates are gathered from coauthors in this paper and college students. Overall, we find that these prompts produce highly similar explanation outputs across models. The generated explanations largely preserve the same core definition and supplementary narratives, with differences primarily in surface-level phrasing (e.g., synonyms and sentence ordering). Therefore, we adopt a single standardized prompt in the benchmark to ensure consistency and emulate realistic Human-LLM interactions. Given a concept `[concept]`, we evaluate the following prompt templates:

- **P1:** *"Tell me about [concept]. Provide a clear explanation suitable for a layperson."*
- **P2:** *"What does [concept] mean? Please explain it in simple terms for a general audience."*
- **P3:** *"Explain [concept] in plain language for someone without background knowledge."*
- **P4:** *"In a few sentences, explain [concept] to a non-expert."*
- **P5:** *"Explain [concept] for a layperson."*
- **P6:** *"I am not familiar with [concept], please explain to me."*
- **P7:** *"I am stupid, please explain [concept] to me."*
- **P8:** *"My friend does not understand [concept], please explain to him/her."*

Across models and domains, we observe the following consistent patterns when switching between different prompt templates: (1). The primary definition and interpretation of the concept are typically unchanged. (2) Differences are mainly at the lexical level and rather than substantive content changes. (3) Most importantly, the relative evaluation results under our evaluation framework remain consistent across these prompt variants. Based on these findings, we use the standardized prompt (P1) for all scientific concept explanation generation.

D Prompt Templates of Explanation Evaluation

This section details the prompt templates we used for the proposed principle-guided evaluation framework in this study:

- **Figure 8**: Prompt template for Direct Scoring.
- **Figure 9**: Prompt template for Comparative Ranking.

E Case Studies

This section illustrates qualitative case studies of a mental health concept explanations (i.e., *Confirmation Bias*) and their corresponding direct scoring and comparative ranking results:

- **Figure 10**: Examples of concept explanations of *confirmation bias* generated by ten popular API-based and open-source LLMs.
- **Figure 11**: Examples of concept explanations of *confirmation bias* generated by three lightweight aligned LLMs.
- **Figure 12**: Assessment results on GPT-4.1-Mini's concept explanation of *confirmation bias* using Gemini-2.5-Pro-as-a-evaluator.

PROMPT TEMPLATES:

Direct Scoring Prompt.

Role: You are an expert evaluator assessing explanations of concepts for a layperson without sufficient knowledge of the field. Your task is to assess the quality of the explanation based on five key principles.

*Concept: [**Concept**]; Explanation: [**Explanation**]; Reference Definition: [**Definition**]*

Evaluation Principles:

- *Accuracy: Analyze the "Explanation" to identify the sentence(s) that define the concept. The "Explanation" provides a faithful and accurate definition compared to the "Reference Definition".*
- *Accessibility: It uses simple, everyday words and avoids technical jargon, acronyms, or academic phrasing.*
- *Conciseness: It is direct and to the point, delivering valuable information without unnecessary verbosity that could cause the user to lose interest.*
- *Demonstration: It uses analogies, stories, or real-world examples that connect the concept to experiences the layperson would understand.*
- *Utility: It provides concrete, practical steps or suggestions that the user can realistically implement in their life.*

Scoring Guide (1-5 for each principle):

- *5 = Excellent - Fully meets the principle*
- *4 = Good - Mostly meets the principle with minor shortcomings*
- *3 = Adequate - Meets basic requirements but could be improved*
- *2 = Poor - Significant shortcomings in meeting the principle*
- *1 = Very Poor - Fails to meet the principle*

Output Format:

- *Accuracy: <score>*
- *Accessibility: <score>*
- *Conciseness: <score>*
- *Demonstration: <score>*
- *Utility: <score>*

Provide only the scores in the "Output Format" with no additional text or explanation.

Figure 8: Prompt Templates for Direct Scoring.

PROMPT TEMPLATES:

Comparative Ranking Prompt.

Role: You are an expert evaluator assessing explanations of concepts for a layperson without sufficient knowledge of the field. Your task is to compare the quality of two explanations based on five key principles.

*Concept: [**Concept**]; Explanation: [**Explanation**]; Reference Explanation: [**Reference Explanation**]; Reference Definition: [**Definition**]*

Evaluation Principles:

- *Accuracy: Analyze the "Explanation" and "Reference Explanation" to identify the sentence(s) that define the concept. The "Explanation" and "Reference Explanation" provide faithful and accurate definitions compared to the "Reference Definition".*
- *Accessibility: It uses simple, everyday words and avoid technical jargon, acronyms, or academic phrasing.*
- *Conciseness: It is direct and to the point, delivering valuable information without unnecessary verbosity that could cause the user to lose interest.*
- *Demonstration: It uses analogies, stories, or real-world examples that connect the concept to experiences the average person would understand.*
- *Utility: It provides concrete, practical steps or suggestions that the user can realistically implement in their life.*

Comparison Result (Win, Loss for each principle):

- *Win = The quality of "explanation" is better than "Reference Explanation"*
- *Loss = The quality of "Reference Explanation" is better than "Explanation"*

Output Format:

- *Accuracy: <result>*
- *Accessibility: <result>*
- *Conciseness: <result>*
- *Demonstration: <result>*
- *Utility: <result>*

Provide only the results in the "Output Format" with no additional text or explanation.

Figure 9: Prompt Templates for Comparative Ranking.

MENTAL HEALTH CONCEPT: CONFIRMATION BIAS

DeepSeek-V3. *Confirmation bias is when people unconsciously favor information that aligns with what they already believe and disregard anything that challenges those beliefs. This matters because it can trap us in rigid ways of thinking, making it harder to learn, adapt, or see situations objectively. For instance, someone convinced that their diet is healthy might only focus on studies supporting their choices while ignoring research highlighting potential risks, potentially harming their long-term well-being. Recognizing this bias helps us stay open to new perspectives and make better-informed decisions.*

Gemini-2.5-Flash. *Confirmation bias is a common mental shortcut where we tend to notice, interpret, and remember information that supports what we already believe, while often overlooking or dismissing anything that challenges our existing views. Think of it like a mental filter that selectively lets in information that confirms our current opinions. This bias is crucial to understand for our mental well-being because it can trap us in narrow ways of thinking, making it hard to learn new things or change our minds, even when faced with new evidence. For example, if you're convinced a particular diet is the best, you might only read articles praising it and ignore any research that suggests otherwise. This selective attention can prevent us from seeing the full picture, leading to poor decisions, reinforced stereotypes, and a struggle to adapt our perspectives, ultimately limiting our personal growth and our ability to interact effectively and empathically with a diverse world.*

GPT-4.1-Mini. *Confirmation bias is the natural tendency to favor information that supports what we already believe while disregarding anything that challenges those beliefs. This bias matters because it can shape our view of the world in a way that isn't fully accurate, potentially leading to poor decisions and limiting our ability to grow or understand others. For example, a person convinced that a particular food is unhealthy might only notice and remember evidence that agrees with them, ignoring research that says otherwise. Recognizing this bias helps us stay open-minded and make better choices for our mental well-being.*

o4-Mini. *Confirmation bias is our mind's habit of focusing on evidence that supports what we already think and dismissing anything that challenges us. This bias can keep us stuck in wrong assumptions, affect our decisions, and maintain unhelpful beliefs that harm our well-being. For example, if someone believes a coworker dislikes them, they might remember every slight glance and ignore dozens of friendly interactions, strengthening their worry without valid proof.*

LLaMA-3.2-1B-instruct. *Confirmation bias is a cognitive bias that occurs when we selectively seek out and give more weight to information that confirms our existing beliefs, while ignoring or downplaying information that contradicts them. This can lead to a distorted view of reality, as we tend to focus on the evidence that supports our preconceived notions, rather than considering alternative explanations. For example, a person who believes that climate change is a hoax may selectively read news articles that support their view, while ignoring or downplaying articles that contradict their opinion, ultimately reinforcing their existing belief.*

LLaMA-3.2-3B-instruct. *Confirmation bias is like having a favorite sports team - you tend to watch games that support your team's wins and ignore the losses. In real life, this bias can lead to poor decision-making and a narrow perspective. For example, if you're a fan of electric cars, you might only read articles about their benefits and ignore studies that highlight their limitations. By recognizing confirmation bias, you can make more informed decisions and consider a wider range of perspectives.*

Qwen-3-1.7B. *This bias helps us make quick decisions by focusing on what we already believe, but it can lead to poor judgment. For example, if you think ice cream causes cavities, you might ignore studies showing no link. It's a natural part of thinking, but it can cloud our ability to see reality clearly.*

Qwen-3-4B. *Confirmation bias skews our decisions by making us ignore contradictory evidence. For example, if you believe a product is great, you might remember its positives but overlook its flaws, leading to poor choices. It's a mental shortcut that can cloud objectivity, like a chef who only tastes a dish once and assumes it's perfect, ignoring feedback from others.*

Qwen-3-8B. *Confirmation bias can skew our decisions and understanding by making us overlook facts that challenge our views. For example, someone who believes in a certain political candidate might only follow news sources that praise them, missing important criticisms, which can lead to poor judgment.*

Qwen-3-14B. *It's like only listening to news that agrees with your opinions and tuning out the rest. This can lead to poor decisions or misunderstandings because you're not considering all the facts. For example, if you think a certain political candidate is the best, you might only follow news that praises them, even if other sources show flaws. This bias keeps you stuck in your own version of reality, making it harder to learn or change your mind.*

Figure 10: Ten LLMs' explanations of a mental health concept *Confirmation Bias*.

MENTAL HEALTH CONCEPT: CONFIRMATION BIAS

Qwen-3-4B-SFT. *Confirmation bias is when people tend to focus on information that confirms what they already believe and ignore or question the validity of information that challenges their views. This bias matters because it can limit a person's ability to think critically, consider new ideas, and make balanced decisions, which can affect their mental well-being and relationships. For example, if someone believes that a certain type of food causes them allergies, they might only pay attention to any symptoms they feel after eating that food and disregard any times they ate it without issues, reinforcing their original belief.*

Qwen-3-4B-DPO. *It shapes how we process information, often leading to flawed decisions. For instance, someone who only follows news sources matching their political views might miss alternative viewpoints, reinforcing their existing stance despite contradictory facts.*

Qwen-3-4B-ExpDPO. *Confirmation bias is the tendency to search for and focus on information that supports what you already believe, while overlooking evidence that disagrees. It's a normal feature of how brains simplify a complex world. However, in mental health, it can make anxiety, low mood, or low self-esteem feel "proven" even when the full picture is mixed. For example, if you believe you're failing, you may zoom in on mistakes and dismiss compliments or successes, which strengthens the belief and can shape your choices. A practical way to counter confirmation bias is to treat your belief like a hypothesis: deliberately collect evidence on both sides, consider alternative explanations, and ask what you would notice if your belief were wrong. Over time, this balanced approach can reduce emotional intensity and support more accurate, compassionate self-understanding.*

Figure 11: Three lightweight aligned LLMs' explanations of a mental health concept *Confirmation Bias*.

EVALUATION RESULTS OF CONFIRMATION BIAS FROM GEMINI-2.5-PRO:

LLM's prompt for Direct Scoring:

Role: You are an expert evaluator assessing explanations of concepts for a layperson without sufficient knowledge of the field. Your task is to assess the quality of the explanation based on five key principles.

*Concept: [**Confirmation Bias**]*

*Explanation: [**Confirmation bias is the natural tendency to favor information that supports what we already believe while disregarding anything that challenges those beliefs ...**]*

*Reference Definition: [**Confirmation bias is a cognitive heuristic characterized by the systematic tendency of individuals to preferentially seek, interpret, and recall information in a manner that confirms their preexisting beliefs or hypotheses, while giving disproportionately less consideration to alternative possibilities or contradictory evidence.**]*

Evaluation Principles: "..."; Scoring Guide: "..."; Output Format: "..."

Provide only the scores in the "Output Format" with no additional text or explanation.

LLM's response:

Accuracy: <5>; Accessibility: <4>; Conciseness: <4>; Demonstration: <5>; Utility: <5>

LLM prompt for Comparative Ranking:

Role: You are an expert evaluator assessing explanations of concepts for a layperson without sufficient knowledge of the field. Your task is to compare the quality of two explanations based on five key principles.

*Concept: [**Confirmation Bias**]*

*Explanation: [**Confirmation bias is the natural tendency to favor information that supports what we already believe while disregarding anything that challenges those beliefs ...**]*

*Reference Explanation: [**Confirmation bias skews our decisions by making us ignore contradictory evidence. For example, ...**]*

*Reference Definition: [**Confirmation bias is a cognitive heuristic characterized by the systematic tendency of individuals to preferentially seek, interpret, and recall information in a manner that confirms their preexisting beliefs or hypotheses, while giving disproportionately less consideration to alternative possibilities or contradictory evidence.**]*

Evaluation Principles: "..."; Comparison Result: "..."; Output Format: "..."

Provide only the results in the "Output Format" with no additional text or explanation.

LLM response:

Accuracy: <Win>; Accessibility: <Win>; Conciseness: <Win>; Demonstration: <Win>; Utility: <Win>

Figure 12: LLM evaluators' prompt template and results on explanation of *Confirmation Bias* (generated by GPT-4.1-Mini). We evaluate it with Gemini-2.5-Pro using Direct Scoring and Comparative Ranking. The reference example is from Qwen-3-4B.