Uncovering Factor-Level Preference to Improve Human-Model Alignment

Anonymous ACL submission

Abstract

Large Language Models (LLMs) often exhibit tendencies that diverge from human preferences, such as favoring certain writing styles or producing verbose outputs. While crucial for improvement, identifying the factors driving these misalignments remains challenging due to existing evaluation methods' reliance on coarse-grained comparisons and lack of explainability. To address this, we introduce PROFILE, an automated framework to uncover and measure the alignment of factor-level preferences of humans and LLMs. Using PROFILE, we analyze preference alignment across summarization, instruction-following, and document-based question-answering tasks. We find a significant discrepancy: while LLMs show poor factor-level alignment with human preferences when generating texts, they demonstrate strong alignment in evaluation tasks. We demonstrate how leveraging the identified generation-evaluation gap can be used to improve LLM alignment through multiple approaches, including fine-tuning with self-guidance.

1 Introduction

Human preference for a piece of text is inherently multifaceted, influenced by an intricate interplay of factors such as fluency, helpfulness, and conciseness. The relative importance of these factors is not static; it often shifts depending on the specific task and context. For instance, a desirable summary should be concise and to the point, while creative writing might prioritize novelty and an engaging narrative. As large language models (LLMs) generate increasingly human-like text, a critical question arises: do these models truly capture the nuance of these varied human expectations, particularly in how they prioritize these underlying quality factors when generating responses?

This question is particularly relevant given existing research highlighting a discrepancy between

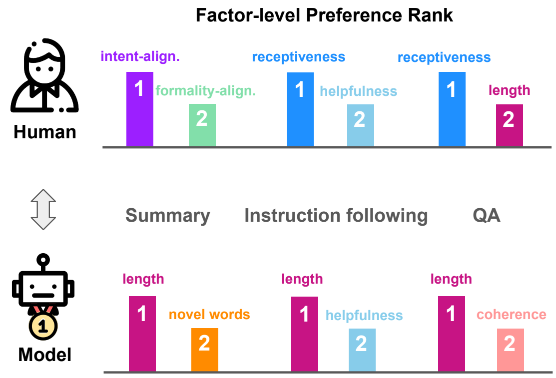


Figure 1: PROFILE uncovers that models exhibit misalignments with human preferences when generating texts. While humans prioritize different quality factors for different tasks, models show consistent bias towards longer output.

LLM's internal encoding (parametric knowledge) and external behavior (Orgad et al., 2024). Notably, recent studies (West et al., 2023; Oh et al., 2024) suggest that a model's ability to generate responses and its ability to discriminate between them are not necessarily aligned. We extend this line of inquiry to investigate whether this discrepancy also manifests at the factor-level of preference alignment. Specifically, do models prioritize individual quality factors consistently with human judgment during generation, and does this prioritization differ when they are asked to discriminate between outputs?

To address these questions, we introduce PROFILE (Probing Factors of Influence for Explainability), an automated framework designed to dissect and quantify how individual factors (e.g., fluency, helpfulness) contribute to overall preference. PROFILE allows us to systematically compare what humans value against what models prioritize, moving beyond surface-level quality scores to uncover deeper, factor-level alignments and misalignments. Using this framework, we seek to answer the following research questions:

1. To what extent do LLMs, during generation,

exhibit factor-level preference alignment with human expectations across various tasks?

2. How does this factor-level alignment compare when the same models perform discrimination tasks (i.e., distinguishing between good and bad responses) versus generation?

3. Can the insights gained from any observed differences in alignment between these settings be leveraged to improve the less aligned aspect?

We conducted experiments using PROFILE to measure LLM alignment with human preferences at a factor level across three key preference alignment tasks—summarization, instruction-following, and document-based QA—evaluating eight prominent LLMs. Our findings reveal a significant misalignment: LLMs often do not prioritize quality factors in line with human expectations during generation. For instance, models frequently exhibit a strong preference for length regardless of the task, whereas human preferences for factors such as conciseness or detail vary contextually (Figure 1).

Interestingly, we observe that these same LLMs demonstrate notably better factor-level alignment when tasked with discriminating between responses. This disparity between generation and discrimination alignment presents an opportunity, and we show that it is indeed possible to leverage the stronger alignment in discrimination to enhance the factor-level preference alignment during generation. Our work underscores the importance of factor-level analysis for a deeper understanding of LLM alignment and offers a pathway toward more genuinely human-aligned generative models.

2 Related Work

Human-AI Preference Alignment. Aligning LLMs with human preferences is a central focus in LLM research, leading to techniques like supervised instruction tuning (Mishra et al., 2021; Wei et al., 2021), RLHF (Ouyang et al., 2022), DPO (Guo et al., 2024), and RLAIF, which utilizes AI-generated feedback (Bai et al., 2022; Lee et al., 2023). However, most studies focus on overall performance (e.g., a response as a whole). While some work has explored using fine-grained human feedback (Dong et al., 2023; Wu et al., 2024), a comprehensive understanding of how granular factors contribute to and differentiate human and model preferences is still lacking. Hu et al. (2023) address this gap by deciphering the factors influencing human preferences. We extend this work by analyzing factor-level preferences across multiple tasks and comparing the driving factors of both humans and models.

Explainable Evaluation of LLMs. Recent research has increasingly emphasized the need for more explainable evaluations of LLMs. For instance, researchers have proposed fine-grained atomic evaluation settings for tasks like fact verification and summarization (Min et al., 2023; Krishna et al., 2023), developed a benchmark for fine-grained holistic evaluation of LLMs on long-form text (Ye et al., 2024), and enhanced evaluation transparency through natural language feedback (Xu et al., 2023). Building on this trend, our work shifts from evaluating individual factors in isolation to analyzing their influence on human preferences and investigating the alignment between human and model judgments regarding the relative importance of these factors.

Furthermore, researchers are actively exploring the potential of LLMs as evaluators. Fu et al. (2024); Madaan et al. (2024); Liu et al. (2023) demonstrate the capacity of large models like GPT-4 to achieve human-like system-level evaluation. However, recent works reveal discrepancies in model performance between generation and evaluation tasks (West et al., 2023; Oh et al., 2024). Inspired by frameworks to meta-evaluate LLM as an evaluator (Zheng et al., 2023; Ribeiro et al., 2020), our work evaluates not only the quality of model-generated text but also the alignment of model preferences in evaluation settings, providing a more comprehensive assessment of LLM capabilities.

3 PROFILE: Framework for Analyzing Human–Model Alignment

PROFILE is a framework designed to uncover and compare the factor-level preferences of humans and language models (Figure 2). It enables:

1. **Interpretation** of overall response quality at a fine-grained factor level. It automatically measures how much each factor (e.g., helpfulness, conciseness, length) contributes to overall preference.

2. **Comparison of alignment between human and model preferences.** PROFILE provides a quantitative score that serves as a metric for human-model alignment in factor level.

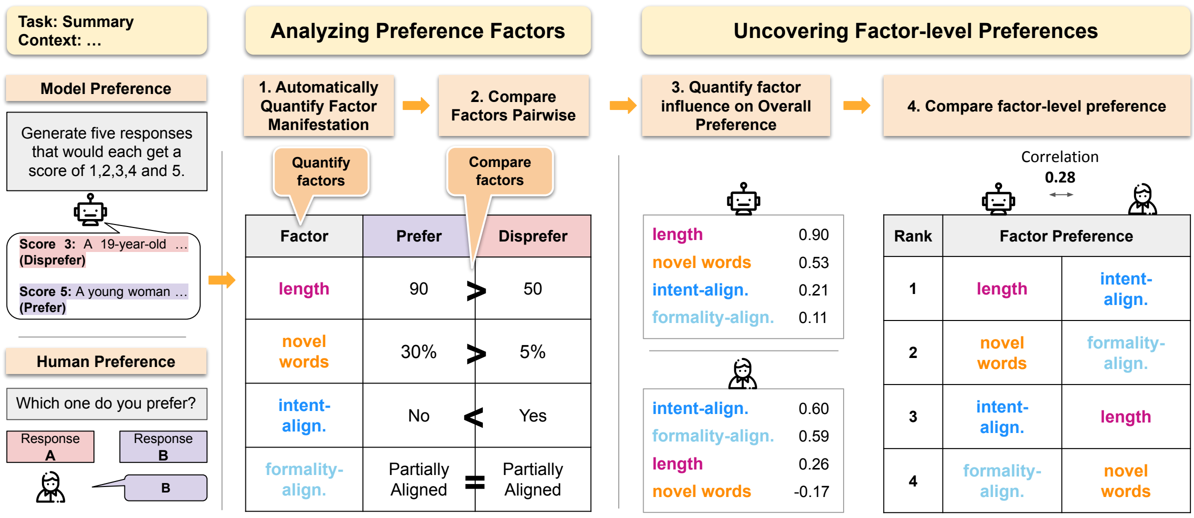


Figure 2: **An Overview of PROFILE pipeline**: (1) extracting factor-level features from responses, (2) comparing them across preferred/dispreferred outputs, and (3) analyzing how each factor contributes to overall preferences in both human and model.

By making these aspects explicit, PROFILE provides a structured way to diagnose and improve alignment between model behavior and human values.

To achieve this, PROFILE follows a three-step process: First, we predefine a set of factors that influence preferences (Table 1). For each response, we automatically quantify the presence or strength of these factors. Next, we compare these factor manifestations between each pair of responses (§ 3.2). Finally, we analyze how these pairwise factor differences correlate with the overall preference labels. This allows us to infer the influence of each factor on overall preference (§ 3.3) and to assess how closely model preferences align with those of humans at the factor level.

3.1 Operational Definitions

We aim to uncover underlying *factor-level* preferences by analyzing observable *response-level* preferences in a pairwise setting, where an agent determines the better of two responses.

Pairwise Preferences. We define the overall response-level pairwise preference function $Pref(r_i, r_j)$ for a given pair of responses $\{r_i, r_j\}$ as:

$$Pref(r_i, r_j) = \begin{cases} 1, & \text{if } r_i \text{ is preferred over } r_j \\ -1, & \text{if } r_j \text{ is preferred over } r_i \\ 0, & \text{if there is a tie} \end{cases}$$

Model Preference in Generation. Models' preference in generation is traditionally defined using log likelihood. $P(x) = \sum_{i=1}^{n} \log P(x_i | x_{<i})$. While this is a direct measure of generation preference, manipulating logits to obtain distinctive outputs presents a technical challenge, and log probabilities are often inaccessible for closed models.

As an alternative, we use score-based prompting[1] as a proxy measure of model preference in generation. In this approach, we instruct the LLM to generate a response *conditioned on achieving* a target quality score ranging from 1 to 5. We then define the model's preference based on these *target input scores*. For instance, if response r_i was generated with a higher target score than response r_j, we define that the model "prefers" r_i in this generation context.

To validate whether the score-based prompting approach effectively approximates the models' intrinsic generation preferences, we conducted an experiment using 100 samples from summarization tasks. Specifically, we prompted open-source models (Llama-3.1-70B and Mixtral) to generate distinct summaries corresponding to target scores ranging from 1 to 5. For each generated summary, we then computed its log probability. We observed a strong Pearson correlation between the target scores and the log probabilities of the generated summaries (Llama-3.1-70B: 0.975; Mixtral: 0.82; see Figure 5 in the Appendix for details).

These results suggest that our scoring mechanism serves as an effective proxy for the models' intrinsic generation preferences, as reflected in their

[1] This approach is inspired by methods used in constructing training dataset for LLM-as-a-judge (Kim et al., 2023).

Factor	Description	Tasks
Receptiveness	Whether the core question of the input has been answered.	*I*, *Q*
Off Focus	The ratio of atomic facts that are not related to the main focus of the input.	*S*, *I*, *Q*
Intent Align.	Whether the intent of the source and output is the same.	*S*
Hallucination	The ratio of atomic facts that are incorrect compared to the original source.	*S*, *I*, *Q*
Source Coverage	The ratio of atomic facts in the source that appear in the output.	*S*
Formality Align.	Whether the formality of the source and output is the same.	*S*
Novel Words	The ratio of words in the output that are not used in the source.	*S*
Length	The number of words used in the output.	*S*, *I*, *Q*
Fluency	The quality of individual sentences of the output.	*S*, *I*, *Q*
Number Of Facts	The number of atomic facts in the output.	*S*, *I*, *Q*
Helpfulness	The ratio of facts that provide additional helpful information.	*I*, *Q*
Misinformation	The ratio of facts in the output that include potentially incorrect or misleading information.	*I*, *Q*
Coherence	Whether all the sentences of the output form a coherent body.	*S*, *I*, *Q*

Table 1: The full taxonomy factors, definitions, and associated tasks (S: Summarization, I: Instruction-following, Q: DocumentQA).

probabilistic outputs.

Thus, model preference in generation, $Pref_{gen}$, is defined by comparing the model-assigned scores $Score(r_i)$ and $Score(r_j)$. $Pref_{gen}(r_i, r_j)$ is 1 if $Score(r_i) > Score(r_j)$, and -1 if $Score(r_i) < Score(r_j)$. In this score-based generation context, we assume distinct scores for different responses, so a tie (0) case for $Pref_{gen}$ is not considered.

3.2 Analyzing Preference Factors

Taxonomy of Preference Factors. To provide a structured framework for analyzing preferences across diverse text generation tasks, we develop a unified taxonomy of fine-grained factors relevant to text quality. This taxonomy categorizes the factors influencing preference alignment between humans and LLMs across text generation tasks. Addressing the lack of a unified framework and inconsistent terminology in existing literature, we consolidate evaluation factors from diverse tasks, including summarization, instruction following, and question answering. For summarization-specific factors, we draw from Fu et al. (2024); Hu et al. (2023); Zhong et al. (2022); Fabbri et al. (2021). For instruction-following and document-based question answering, we incorporate categories from Glaese et al. (2022); Ye et al. (2024); Nakano et al. (2021). The complete taxonomy is detailed in Table 1.

Quantifying Factor Manifestation. We employ several approaches to automatically analyze the manifestation of our factors in responses: (i) **Rule-based**: For straightforward, objective factors, we use deterministic algorithms. Length and Novel Words are extracted this way. (ii) **UniEval-based**: For inherently subjective factors (Fluency and Coherence), we use the well-established UniEval metric (Zhong et al., 2022). UniEval is a learned metric that provides scores of range 0-1 for various aspects of text quality. (iii) **LLM-based**: For factors that rely on objective criteria but require more nuanced judgment, we use GPT-4o with carefully designed prompts. This approach is further divided into "response-based" (Intent Alignment and Formality Alignment) and "atomic-fact-based" (the remaining seven) extraction depending on the level of detail needed for each factor. The specific details of the implementation of each method and validation of LLM-based extractions can be found in Appendix D.

Comparing Factors Pairwise. For each pair of responses, we compare factor manifestation. For each factor f, we define a function M_f to compare factor's manifestation in pairs of responses:

$$M_f(r_i, r_j) = \begin{cases} 1, & \text{if } f \text{ is more manifest in response } r_i \\ -1, & \text{if } f \text{ is more manifest in response } r_j \\ 0, & \text{if } f \text{ is equally manifest in both responses} \end{cases}$$

For example, if r_i is longer than r_j, then $M_{length}(r_i, r_j) = 1$.

3.3 Uncovering Factor-level Preferences

Quantifying Factor Influence. To quantify each factor's influence (factor score), we analyze the concordance between response-level preferences $Pref(r_i, r_j)$ and factor manifestation $M_f(r_i, r_j)$ across response pairs. We use τ_{14}, a variation of Kendall's correlation proposed by Macháček and Bojar (2014), which is particularly well-suited for handling ties in our analysis setting, where ties arise in only one of the comparison sets used for calculating Kendall's τ.

The metric is defined as:

$$\tau_{14}(f) = \frac{|C_f| - |D_f|}{|C_f| + |D_f| + |T_f|,} \quad (1)$$

where C_f is the count of concordant pairs (preference and factor manifestation agree), D_f is the count of discordant pairs (preference and factor manifestation disagree), and T_f represents ties. In our setting, since models don't generate responses with identical scores, T_f only counts factor-level ties ($M_f(r_i, r_j) = 0$) This pairwise comparison reveals how the differences in factor manifestations relate to differences in preference between the two responses.

For instance, consider the factor M_{length}, which measures response length. If response r_1 is longer than r_2 ($M_{length}(r_1, r_2) = 1$) and the model prefers r_1 ($Pref(r_1, r_2) = 1$), this pair is classified as concordant. Conversely, if the model prefers the shorter r_1, the pair is discordant. Evaluating all pairs, a positive factor score indicates a positive influence of the factor, a negative score indicates a negative influence, and a score close to zero implies minimal influence. The magnitude of the score reflects the strength of this influence.

Comparing Human and Model Preferences. Finally, we evaluate *factor-level preference* alignment by comparing human and model factor rankings. We use Spearman's ρ, Kendall's τ [2], and Pearson's r coefficients to quantify the correlation between these rankings, providing a measure of how well the model's factor priorities align with human values.

4 Uncovering Factor-Level Preference of LLMs

In this section, we analyze how models and humans differ in their factor-level preferences during text generation. Using human preference datasets across summarization, instruction-following, and QA tasks, we apply PROFILE to model-generated responses and compare the relative importance of each factor with human judgments.

4.1 Experimental Setting

Tasks. We use human preference alignment data publicly available. Among them we choose: (i) Reddit TL;DR (Stiennon et al., 2020), which includes human ratings of summaries across multiple evaluation dimensions; (ii) StanfordHumanPreference-2 (SHP-2) (Ethayarajh et al., 2022), focusing on human preferences over responses in the "`reddit/askacademia`" domain; and (iii) OpenAI WebGPT (Nakano et al., 2021), which compares model-generated answers on the ELI5 subreddit based on factual accuracy and usefulness [3]. We refer to the tasks for each dataset as summarization, instruction-following, and document-based QA tasks in this paper. We exclude pairs with human Tie ratings in all three datasets, as our analysis focuses on cases with clear preference distinctions.

Models. For our experiments, we utilize both open-source and proprietary LLMs. Open-source models include LLaMA 3.1 70B (Dubey et al., 2024), Mixtral 8x7B Instruct v0.1 (Jiang et al., 2024), and three TÜLU v2.5 models (Ivison et al., 2024) (TÜLU v2.5 + PPO 13B (13B RM), TÜLU v2.5 + PPO 13B (70B RM), and TÜLU v2.5 + DPO 13B). Proprietary models include Gemini 1.5 Flash (Reid et al., 2024), GPT-4o (OpenAI, 2024), and GPT-3.5. From here on, we refer to Gemini 1.5 Flash as Gemini 1.5, Mixtral 8x7B Instruct v0.1 as Mixtral, TÜLU v2.5 models as Tulu 2.5 + {alignment training strategy}. Detailed descriptions of the datasets and models can be found in Appendix C.2.

Experimental Setup. For each task, models generate a response that would receive a score of 1-5. The specific prompts we used can be found in Appendix E. Additionally, we find that responses generated with score 5 strongly align with those from direct, unconstrained generation (see Table 13), suggesting the generalizability of our experimental setting.

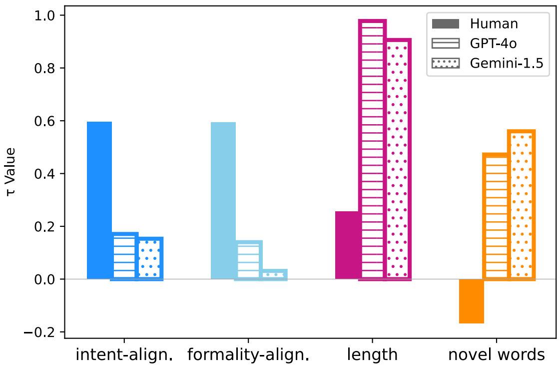
4.2 Factor-level Alignment in Generations

	τ	ρ	r
Mixtral	0.200	0.297	0.069
Tulu 2.5 + PPO (13B RM)	-0.156	-0.164	-0.189
Tulu 2.5 + PPO (70B RM)	0.111	0.200	-0.015
LLaMA 3.1 70B	0.111	0.248	0.213
Gemini 1.5	**0.289**	**0.394**	**0.171**
GPT-4o	0.156	0.297	0.155

Table 2: Factor-level preference alignment (τ, ρ, r) between model and human in the generation setting for the summarization task.

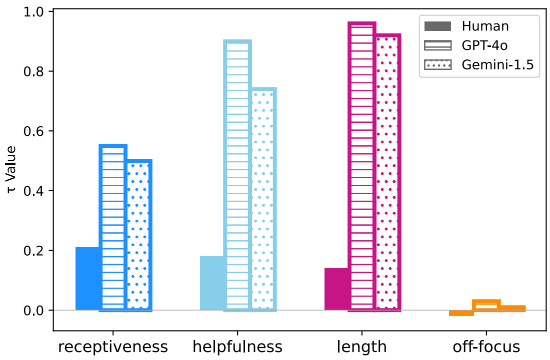
[2] We use Kendall's τ_b (Kendall, 1945) as the default.

[3] Our framework can also be applied to other tasks. We provide guidelines for applying it to different tasks, with an example of a mathematical reasoning task in the Appendix E.2.



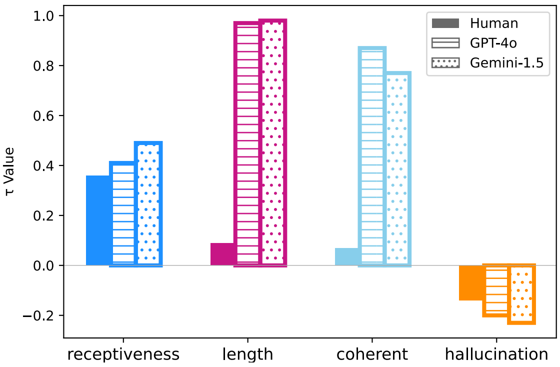
Rank	Human	GPT-4o	Gemini-1.5
1	intent-align.	length	length
2	formality-align.	# facts	# facts
3	# facts	src. coverage	src. coverage
4	src. coverage	novel words	coherence
5	length	intent-align.	novel words
6	coherence	coherence	intent-align.
7	off-focus	hallucination	formality-align.
8	hallucination	formality-align.	hallucination
9	fluency	off-focus	fluency
10	novel words	fluency	off-focus

(a) Summarization



Rank	Human	GPT-4o	Gemini-1.5
1	receptiveness	length	length
2	helpfulness	helpfulness	helpfulness
3	# facts	# facts	# facts
4	length	coherence	coherence
5	fluency	receptiveness	receptiveness
6	coherence	fluency	hallucination
7	misinformation	off-focus	misinformation
8	hallucination	misinformation	off-focus
9	off-focus	hallucination	fluency

(b) Instruction-following



Rank	Human	GPT-4o	Gemini-1.5
1	receptiveness	length	length
2	length	coherence	coherence
3	helpfulness	receptiveness	receptiveness
4	# facts	helpfulness	helpfulness
5	coherence	fluency	# facts
6	fluency	novel word	novel word
7	misinformation	# facts	misinformation
8	off-focus	off-focus	fluency
9	novel word	misinformation	hallucination
10	hallucination	hallucination	off-focus

(c) Document QA

Figure 3: **PROFILE uncovers the factor-level preferences of humans and models.** Figure illustrates the comparison of factor-level preference alignment between humans, GPT-4o, and Gemini-1.5 in generation across three tasks: (a) Summarization, (b) Instruction-following, and (c) Document QA task. The left bar graphs display *factor scores* (τ_{14}) for selected factors. The right tables show the rankings of all factors for each task. Notably, both models consistently rank 'length' as the top factor across tasks, while human preferences vary by task.

PROFILE uncovers the factor-level preferences of humans and models. PROFILE enables fine-grained analysis of preference alignment by breaking down overall judgments into interpretable factor-level scores. This allows us to identify not only how models and humans differ in ranking specific factors (Figure 3), but also to quantify their alignment using correlation metrics (Table 2). Through this, PROFILE reveals consistent patterns of agreement and misalignment that would be obscured by aggregate quality scores alone.

Human and model preferences consistently misalign at the factor level across tasks. While humans' most preferred factors vary by task, models consistently prioritize length across all tasks, suggesting models associate better quality with longer outputs. In both instruction-following tasks (Figure 3b) and document-based QA (Figure 3c), humans prioritize Receptiveness and Helpfulness. Although these two factors are also highly ranked for the models, the models always prioritize Length as the most important factor.

The misalignment pattern is particularly problematic in summarization tasks. Humans prioritize IntentAlignment, FormalityAlignment, and SourceCoverage while penalizing the inclusion of words not in the original post, indicating the importance of maintaining the original content and style. In contrast, models consistently prefer longer summaries with new words (Table 7). A full list of factor scores of all models across three tasks is available in the Appendix (Table 9 - 11).

To quantify this misalignment, we measure *factor-level preference alignment* (τ). The left Generation column in Table 2 shows that even the best-performing model (Gemini 1.5) only achieves a 0.289 τ correlation with human preferences in summarization. Similar low correlations are observed in the other two tasks (Table 12). This low correlation highlights the limitations of current models in capturing the granular aspects comprising human preference.

Qualitative analysis demonstrates how PROFILE explains the observed misalignment. In a Reddit post above, GPT-4o's score 5 summary is longer and includes more facts than its score 3 summary, yet the shorter summary is human-preferred. The higher-scored model summary includes irrelevant details like "Midwest hometown" and "new to Reddit," demonstrating the model's tendency to prioritize information quantity over relevance. Full examples are in Appendix A.2.

GPT-4o Generation Sample

Post: Good Morning/Afternoon r/advice, Never posted on Reddit before at all, but I figured (based on the overall reliability of you nice individuals) (...)
Score 5 [length: 93, # facts: 10, src. coverage: 0.389]: A Reddit user **recently moved back to their Midwest hometown** and, while setting up utilities for their new place, discovered they owe \$500 in gas bills from a college house they lived in until 2012. (...)
Score 3 [length: 61, # facts: 9, src. coverage: 0.44]: A Reddit user seeks advice after discovering they owe \$500 in gas bills from a college house they left in 2012. (...) **(Human Preferred Output)**

5 Achieving Better Alignment Using the Model as a Discriminator

Building on the observations from prior works that LLMs with the same parametric knowledge may show different external behaviors in different settings, such as in discriminatory tasks (West et al., 2023) or evaluation tasks (Oh et al., 2024), this section explores whether LLMs exhibit different factor-level preference alignment when performing this task, and, if so, whether we can leverage any observed improvements in alignment to guide better generation.

	Gen.	Eval.	
	τ	τ	*Agree. (%)*
Mixtral	0.200	0.244	0.526
Tulu 2.5 + PPO (13B RM)	-0.156	0.511	0.516
Tulu 2.5 + PPO (70B RM)	0.111	0.644	0.520
LLaMA 3.1 70B	0.111	0.733	0.705
Gemini 1.5	**0.289**	<u>0.778</u>	<u>0.721</u>
GPT-4o	0.156	**0.822**	**0.784**

Table 3: Kendall's τ correlation in generation and evaluation settings, and evaluation agreement rate (%) for the summarization task.

LLMs are more aligned in evaluation than in generation. Models demonstrate significantly stronger alignment with human preferences in evaluation tasks compared to generation. Table 3 demonstrates this by showing *factor-level preference alignment* of human and model, measured using Kendall τ is consistently higher in the evaluation setting across all models. For instance, GPT-4o exhibits the highest alignment in evaluation (τ: 0.82) but much lower alignment in generation (τ: 0.16).

These findings raise a natural question: **can we leverage this alignment not only to evaluate but to *actively improve generation quality*?**

	GPT-4o		LLaMA 3.1 70B		Tulu 2.5 + PPO (70B RM)	
	τ_G	τ_H	τ_G	τ_H	τ_G	τ_H
Baseline$_A$	-0.24	−0.07	−0.20	−0.29	−0.29	−0.29
Baseline$_B$	-0.29	−0.29	−0.42	−0.42	−0.24	−0.24
GPT-4o feedback	**0.36**	**0.45**	**0.29**	**0.20**	**0.16**	**0.16**

Table 4: Factor-level alignment (τ) between improvements made by different generators (GPT-4o, LLaMA 3.1 70B, Tulu 2.5 + PPO (70B RM)) and factor-level preferences from GPT-4o (evaluation) and human. τ_G and τ_H indicate alignment with GPT-4o and human preferences respectively. Higher values show stronger alignment.

	τ	ρ	r
Tulu 2.5 w/o SFT	0.111	0.2	-0.015
Tulu 2.5 self-SFT	**0.156**	**0.297**	**0.028**

Table 5: Factor-level preference correlations between humans and Tulu 2.5 (70B RM) with and without supervised fine-tuning from self-evaluation (self-SFT).

To answer this, we explore whether LLMs' discriminative behavior-either through self-evaluation or external feedback-can guide models toward better factor-level alignment during generation. Below, we test both self-refinement through supervised fine-tuning and feedback-driven generation, grounded in the observed gap between models' evaluation and generation preferences.

Gen-Eval Gap Explains Self-refinement's Effectiveness. We investigate whether supervised fine-tuning (SFT) with self-evaluation can improve preference alignment in generation. Using TULU 2.5 (70B RM), we generate 1-5 score summaries, then use the same model to pairwise evaluate and re-rank these summaries based on win rate. The generator is then SFT-trained on 4,000 such examples and tested on 500 unseen examples. The input is an instruction to generate summaries of scores 1-5 given a post, and output labels are the re-ranked summaries of score 1-5. Table 5 shows the SFT-trained model achieves significantly improved alignment compared to the original TULU model, reaching performance comparable to GPT-4o (Table 2).

Leveraging Evaluation for Better Alignment in Generation. We explore whether explicit feedback from a strong evaluator can improve summary generation. A generator model produces two initial summaries per post, and an evaluator selects the preferred one (or tie) and provides a justification. The generator then uses this feedback to generate an improved summary. Using GPT-4o as the evaluator, we compare a feedback-driven approach with two baselines: (1) Baseline$_A$, where the generator produces one improved summary from both initial summaries *without* feedback; and (2) Baseline$_B$, where the generator produces two improved summaries *without* feedback, each based on one initial summary. These baselines represent typical improvement scenarios relying on implicit self-critique. Experiments are conducted on 100 Reddit TL;DR samples with three generators (GPT-4o, LLaMA 3.1 70B, and Tulu 2.5 + PPO.

Table 4 shows that incorporating evaluator feedback leads to improved alignment, correlating positively with both GPT-4o and human judgments across all generators. In contrast, the baselines, which rely on re-generation without explicit feedback, show negative correlations, indicating a divergence from the desired preferences. Manual analysis of 30 samples confirms that evaluator feedback emphasizes higher-ranked factors in the evaluator's preferences (with the exception of Formality Alignment; see Appendix F.2.3). These results demonstrate the effectiveness of leveraging external evaluation feedback for enhancing generation alignment. See Appendix F.2.1 for prompt and metric details.

6 Conclusion

We introduce PROFILE, a framework for granular factor level analysis of LLM alignment with human preferences. Our analysis using PROFILE reveals that LLMs tend to over-prioritize factors like output length, misaligning human preferences during generation. However, these models exhibit stronger alignment in evaluation tasks, indicating the potential for leveraging evaluative insights to improve generative alignment. PROFILE facilitates a nuanced understanding of the alignment gaps and mismatches between human and model preferences. These insights underscore the necessity for more sophisticated, factor-level alignment strategies that can guide the development of LLMs to better align with human expectations, ultimately fostering more reliable aligned AI systems.

7 Limitations

This study has several limitations. First, the preference datasets used may not fully represent the entire spectrum of human preferences. Second, due to budget constraints, human evaluations of model outputs were conducted on a limited scale, with a restricted number of participants, and only on one task. Furthermore, this study represents a preliminary exploration into methods for achieving better alignment, highlighting the potential of various techniques to enhance generation and evaluation. Extensive studies are required to thoroughly assess the efficacy and generalizability of these methods. While this study focuses on post-hoc correction methods, future research should investigate how to incorporate the identified preference factors as signals during the training stage. Additionally, exploring how to embed these signals within datasets used for preference optimization represents a promising direction for future work.

8 Ethics Statement

Our research relies on established benchmarks and models, and does not involve the development of new data, methodologies, or models that pose significant risks of harm. The scope of our experiments is limited to analyzing existing resources, with a focus on model performance. Human studies conducted within this work adhere to relevant IRB exemptions, and we ensure fair treatment of all participants. Our work is mainly focused on performance evaluation, we recognize that it does not specifically address concerns such as bias or harmful content.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. 2023. DecipherPref: Analyzing influential factors in human preference judgments via GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8344–8357, Singapore. Association for Computational Linguistics.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *arXiv preprint arXiv:2406.09279*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Jungsoo Won, Dongha Lee, and

Jinyoung Yeo. 2024. Evaluating robustness of reward models for mathematical reasoning. *Preprint*, arXiv:2410.01729.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*.

Juhyun Oh, Eunsu Kim, Inha Cha, and Alice Oh. 2024. The generative AI paradox in evaluation: "what it can solve, it may not evaluate". In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 248–257, St. Julian's, Malta. Association for Computational Linguistics.

OpenAI. 2024. Hello, gpt-4 turbo. https://openai.com/index/hello-gpt-4o/.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. 2023. The generative ai paradox:"what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Appendix

A Human Evaluation of Model Generations

We collect human preference data via Amazon Mechanical Turk (MTurk) for 30 posts and 6 models. For each post, three summary pairs—selected from five model-generated summaries (scored 1 to 5)—are presented to three annotators. Annotators, restricted to US-based workers with a 95% and HIT approval rate and over 5,000 approved HITs are recruited. The MTurk task description clearly explained the study's purpose and data usage. As shown in Figure 4, we provide detailed instructions about the experiment through MTurk, and participants who consented then participated in the study.

For the main experiment, we gave annotators the following instructions: *"A good summary is a shorter piece of text that captures the essence of the original. It aims to accomplish the same purpose and convey the same key information as the original post. Please choose the better summary, A or B."* Each annotation is compensated at $0.25. This process yields 1,620 annotations (30 posts * 6 models * 3 pairs/model/post * 3 annotations/pair).

For each model and corresponding summary pair, we calculate an agreement rate, which indicates the percentage of pairs where at least two out of three annotators prefer the summary with the higher score assigned by the model. Each model is evaluated on 90 summary pairs, derived from 30 posts with three pairs per post.

A.1 Examples of Human-Model Misalignment

A higher score summary generated by GPT-4o as shown by the factor-level analysis (in magenta), is longer and includes details not necessary to understand the essence of the original post.

A.2 Human Evaluation Results of Model Generations

The agreement rates for each model are summarized in Table 6. Despite low factor-level preference alignment (τ), overall agreement rates range from 56% to 75%. This suggests that binary evaluations, where annotators choose the better summary based on overall quality, can mask factor-level misalignments such as preferences for summary length. Since models consistently favor longer responses, continued reliance on this setup to generate human-annotated training data may reinforce the tendency towards verbose answers, aligning with human preferences superficially but missing finer-grained alignment on factors like length or conciseness.

GPT-4o misaligned examples

Post: Good Morning/Afternoon r/advice, Never posted on Reddit before at all, but I figured (based on the overall reliability of you nice individuals) that now would be a good time to start.

About a year ago I moved away from my Midwest town to the west coast for work, and have just recently moved back. I was calling the Utility providers in my area to set up the gas/water/elec for my new place, but I was informed that gas could not be set up due to owing a past amount.

(...)

My question is, am I really on the hook for these fees and what recourse could I possibly have? I cannot afford to pay it, but I also don't think that it should be on me TO pay it. Any advice is truly appreciated as I have not had any experience with this type of thing before.

GPT-4o's score 5 generation [length: 93, # facts: 10, src. coverage: 0.389]: A Reddit user recently moved back to their Midwest hometown and, while setting up utilities for their new place, discovered they owe $500 in gas bills from a college house they lived in until 2012. The user didn't cancel the service when they moved out, leading to the bill accruing over time. They are looking for advice on whether they are responsible for this debt and what recourse they might have, as they are unable to afford the payment. The user is new to Reddit and appreciates any guidance on handling this situation.

GPT-4o's score 3 generation [length: 61, # facts: 9, src. coverage: 0.44]: A Reddit user seeks advice after discovering they owe $500 in gas bills from a college house they left in 2012. They didn't cancel the service when they moved out, and the bill continued to accrue. They want to know if they are responsible for paying it and what actions they can take, as they can't afford to pay the bill. **(Human Preferred Output)**

	τ	*Agree.(%)*
Gemini 1.5	**0.289**	0.689
Mixtral	<u>0.2</u>	0.678
GPT-4o	0.156	<u>0.722</u>
LLaMA 3.1 70B	0.111	**0.756**
Tulu 2.5 + PPO (70B RM)	0.111	0.567

Table 6: Factor-level preference alignment (τ) and response-level agreement with human preferences in the summarization Task.

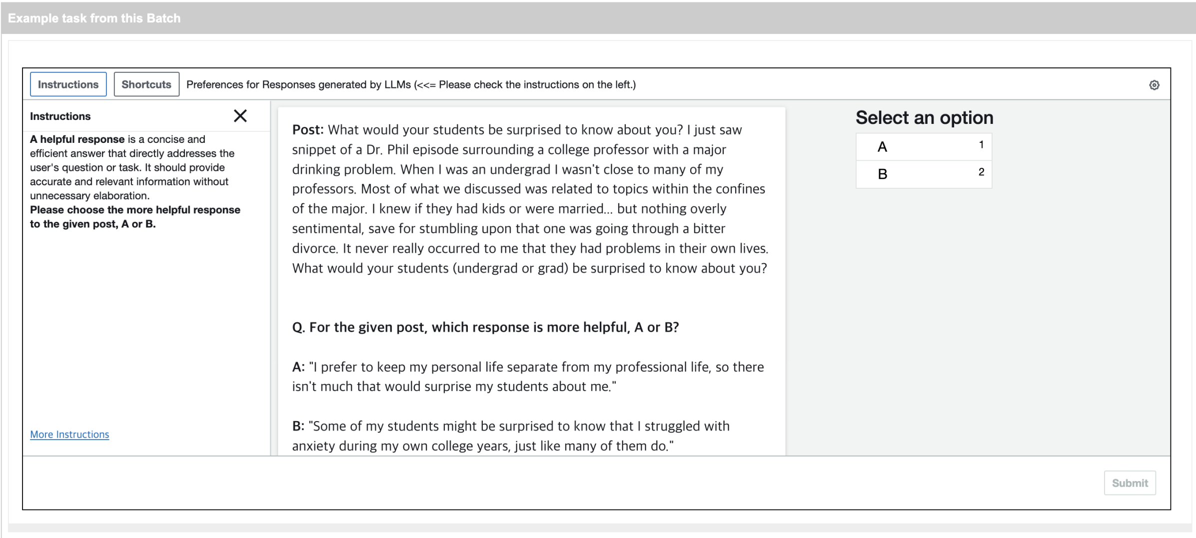


Figure 4: A screenshot of a sample summary with preference annotations.

B Additional Examples of Model Generations and Evaluations

B.1 Generation

Table 7 shows summaries of scores 1-5 that GPT-4o generated. We can see a clear trend of higher score summaries being longer, with more information density. The information additional to lower-scoring summaries is the ages of the characters, which is not essential in understanding the original post. This is a randomly selected example and a typical model-generated summary case.

Additionally, we analyzed cases where model-generated outputs received higher scores but are not preferred by human annotators (based on majority votes). The human-model misalignment is more pronounced when comparing summaries with larger score gaps assigned by the models. We provide additional examples of such divergence for GPT-4o, LLaMA 3.1 70B and Tulu 2.5 + PPO (13B RM). We find that when the length and information gap is significant, humans favor summaries that are shorter and less information-dense, as model-generated summaries with lower scores (1-2) generally convey the main intention of the post as well. When factual content is similar, humans tend to favor more informal responses (e.g., “Guy” over “An 18-year-old male”, and “gotten back together” over “renewed romantic connection”), which aligns with our factor-level findings indicating higher formality-alignment for human-preferred outputs, since Reddit posts are mostly informal.

Warning: Examples below contain content that may be offensive or upsetting.

GPT-4o misaligned examples

GPT-4o’s score 4 generation: An 18-year-old male in a long-distance relationship with his 17-year-old girlfriend told her he loves her but wants an open relationship. The girlfriend, who has been with four other boys before him, reacted very negatively, telling him to “FUCK OFF" and that she never wants to speak to him again. He is seeking advice on how to make the situation easier for her and whether he can maintain their relationship while exploring other people.

GPT-4o’s score 1 generation: Guy told his girlfriend he wants an open relationship. She got mad and told him to go away. He wants advice. **(Human Preferred Output)**

Llama-3.1-70b misaligned examples

Llama’s score 4 generation: I’m in a relationship with an 18-year-old girl who’s a self-described Tsundere. We share similar interests, but her mood swings are causing problems. She can be affectionate and loving, but also violent and aggressive, lashing out at me over minor issues. For example, she threw shoes at me for a joke and almost stabbed me with a knife over a trivial matter. I’m scared and unsure how to navigate these outbursts.

Llama’s score 2 generation: I’m dating a girl who’s similar to me, but she’s really violent and has mood swings. She gets angry over small things and has even attacked me. I’m scared and need help. **(Human Preferred Output)**

Score 1	Score 2	Score 3	Score 4	Score 5
The poster and their fiance have scabies and are treating it. Their roommate refuses to get treated or clean his things, risking reinfestation. The poster is considering asking him to move out.	A couple living with a roommate got scabies and started treatment. The roommate, who has no symptoms, refuses to treat himself or his belongings. The poster is worried about reinfestation and may ask the roommate to move out since he isn't on the lease.	A 19-year-old woman and her 20-year-old fiance contracted scabies after moving into a new apartment with a mutual friend. Both are undergoing treatment, but the roommate refuses to get treated or clean his belongings. The poster is concerned about reinfestation and is contemplating asking the roommate to move out, as he is not on the lease.	A 19-year-old woman and her 20-year-old fiance recently moved into a new apartment with a mutual friend and contracted scabies. Despite both of them undergoing treatment, their roommate refuses to take any preventive measures or treat himself, even though he has not shown symptoms yet. The poster is worried that this will lead to a reinfestation and is considering asking the roommate to move out since he is not on the lease.	A young woman (19) and her fiance (20) recently moved into a new apartment with a mutual friend and unfortunately contracted scabies. Both have started treatment after a doctor's confirmation. However, their 20-year-old roommate, who has not shown symptoms, refuses to undergo treatment or clean his clothes and linen. This refusal has caused concern for the woman, as she fears reinfestation and is seriously considering asking the roommate to move out, especially since he is not on the lease.

Table 7: An example of GPT-4-generated summaries of scores 1-5.

Tulu 2.5 + PPO (13B RM) misaligned examples

Tulu's score 4 generation: The OP discovered a potentially incriminating text exchange between their SO and her ex's sister, suggesting a renewed romantic connection.

Tulu's score 1 generation: The OP read a suspicious text on their SO's phone, sent from her ex's sister, that suggested they might have gotten back together. **(Human Preferred Output)**

B.2 Evaluation

We provide examples where the model evaluations align with human preferences, even if the chosen option contains less facts or is shorter. In the first example, where both GPT-4o and LLaMA 3.1 70B correctly chose human-preferred summary, while the chosen summary is shorter, it more accurately reflects the key issue in the original post by mentioning the writer's economic status. In the second example, the GPT-4o chosen summary is more clearly reflecting the content in post over the other option which analogically describes the main idea of the post.

GPT-4o & LLaMA aligned examples

Post: Yesterday, I accidentally dropped my Motorola Atrix 2 and the screen cracked really badly. My phone is still fully functional, but it's a bit difficult to see what I'm doing when I'm texting or web browsing, etc. Anyway, I stupidly didn't buy insurance for my phone and I'm not eligible for an upgrade until next May! AT&T offers some options as far as getting a no-commitment phone at a slight discount, but spending $300-$600 for a new phone isn't really in the budget right now.
(...)
I found a couple websites that will repair your phone if you send it in. [Doctor Quick Fix] will do it for $110 and I'm still waiting on a quote from [CPR](So my question is, have any of you used this company, or know anyone who has used it? Should I trust these companies? Do you have any recommendations? What should I do to get my phone fixed?

Summary A: Dropped my phone, they said they won't repair phones that have been physically abused. Looking for suggestions on cell phone repair companies, if any, and what I should do to get my phone fixed.

Summary B: I dropped my phone, cracking the screen. I can't afford to buy a full price phone, so should I try the above repair companies? What should I do? **(Human Preferred Output)**

GPT-4o aligned & LLaMA misaligned examples

Post: I got a letter in the mail saying I've been passed up for being hired for my dream job. I wanted this job for 10 damn years and now it's over. I've trained my body, mind, and soul for this job and just through a simple letter, I've been removed from that process. I was in good standing with getting hired. Passed everything with flying colors.
(...)
Now what? Am I to live with my parents the rest of my life? Am I to never get my dream car? Am I to just keep my job where I only get paid minimum wage while I make the company tens of thousands? I don't know what to do. I mean my second dream job would be to work with penguins, but I don't think that's possible for me. Anyone have any advice for me? What should I do?

Summary A: I followed the yellow brick road for half my life and ended up at a complete dead end and I can't turn around to go back.

Summary B: Got passed up for a dream job. Now what the hell are I supposed to do with my life that doesn't include my dream job? **(Human Preferred Output)**

C Experimental Setting

C.1 Tasks

We examine three publicly available datasets of pairwise human judgments commonly used in preference optimization methods like RLHF and DPO training: **Reddit TL;DR** We analyze the dataset released by OpenAI (Stiennon et al., 2020), which includes human ratings of summaries across multiple axes (referred to as "`axis evaluations`"). Higher scores indicate human preference across multiple evaluation dimensions. **StanfordHumanPreference-2 (SHP-2)** (Ethayarajh et al., 2022), focuses on capturing human preferences over responses to questions and instructions, prioritizing helpfulness. Higher scores indicate a more helpful response. For this study, we use responses from the "`reddit/askacademia`" domain. **OpenAI WebGPT** This dataset (Nakano et al., 2021), addresses the task of generating answers to questions from the ELI5 (*"Explain Like I'm Five"*) subreddit. Human annotations compare two model-generated answers based on factual accuracy and overall usefulness. We exclude pairs with Tie ratings in all three datasets, as our analysis focuses on cases with clear preference distinctions.

C.2 Models

Our study focuses on the most advanced and widely-used generative models currently accessible, encompassing both proprietary and open-source options. For open-source models, we include LLaMA 3.1 70B (Dubey et al., 2024)[4], Mixtral 8x7B Instruct v0.1 (Jiang et al., 2024), three TÜLU 2.5 Models (Ivison et al., 2024)—TÜLU 2.5 + PPO 13B (13B RM) [5], TÜLU 2.5 + PPO 13B (70B RM) [6], and TÜLU 2.5 + DPO 13B [7]. For proprietary models, we use Gemini 1.5 Flash (Reid et al., 2024), GPT-4o (OpenAI, 2024) [8], and GPT-3.5 [9]. We set the parameters for all models to: temperature = 0.6, top_p = 0.9, and max_tokens = 1024. 4 Quadro RTX 8000 48GB were used with CUDA version 12.4 when running TULU Models.

We used autotrain library [10] for supervised fine-tuning TULU model in experiments in § 5. The parameters for fine-tuning are as follows: block_size: 2048, model_max_length: 4096, epochs: 2, batch_size: 1, lr: 1e-5, peft: true, quantization: int4, target_modules: all-linear, padding: right, optimizer: paged_adamw_8bit, scheduler: linear, gradient_accumulation: 8, mixed_precision: bf16, merge_adapter: true

D PROFILE

D.1 Validation

Figure 5 shows the distribution of Pearson correlations over 100 samples for both LLaMA-3.1-70B and Mixtral.

We find that the correlation of most samples are concentrated between 0.85 and 1.0, indicating a strong correlation between the target scores in our score-conditioned setting and the models' log probabilities (i.e., their preference for those responses)

D.2 Factor Extraction Methods

Rule-based Extraction We obtain the Length and Novel Words using a rule-based extraction method. First, we calculate the output's length and count the novel words by removing special characters and splitting the text into words. The total word count represents Length. For Novel Words,

[4] Inference for LLaMA was conducted using the Together AI API. `https://www.together.ai/`

[5] We use huggingface allenai/tulu-v2.5-ppo-13b-uf-mean-13b-uf-rm model.

[6] We use huggingface allenai/tulu-v2.5-ppo-13b-uf-mean-70b-uf-rm model.

[7] We use huggingface allenai/tulu-v2.5-dpo-13b-uf-mean model.

[8] We use gpt-4o-2024-05-13 version for all GPT-4o inference.

[9] We use gpt-3.5-turbo-1106 version for all GPT-3.5 inference.

[10] `https://huggingface.co/autotrain`

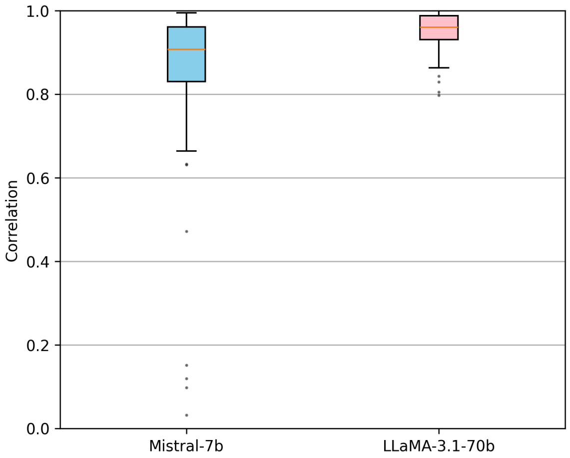


Figure 5: Pearson correlation between target conditioning scores and log probabilities of generated summaries for Mistral-7b and LLaMA-3.1-70b.

we stem both the source text and the model output to create unique sets of stemmed words, then determine the number and proportion of unique words in the output that differ from the source.

LLM-based Extraction The calculations are divided into atomic-fact-level and response-level based on the granularity of the factors.

Atomic-Fact-Level Factors refer to those factors that are evaluated based on the presence or absence of each factor at the atomic fact level. An atomic fact is a short, self-contained piece of information that does not require further explanation and cannot be broken down further (Min et al., 2023). These include the Number Of Facts, Source Coverage, Off Focus, Hallucination, Helpfulness, and Misinformation. The Number Of Facts is determined by counting the total atomic facts, while the remaining factors are calculated as the ratio of relevant atomic facts to the total number of atomic facts.

Response-Level Factors refer to those factors that are evaluated based on the presence or absence of each factor at the response level. These include Receptiveness, Intent Alignment, and Formality Alignment. Formality Alignment is classified into one of three categories: [Aligned/Misaligned/Partially-Aligned], while the other two factors are determined in a binary manner [Yes/No].

The prompts used are provided in D.3. The Source Coverage does not have a separate prompt since it was calculated using the output from the Hallucination (i.e., the ratio of non-hallucinated atomic facts to the total number of atomic facts in the Source Post).

D.3 Prompt Template For LLM-based Factor Extraction

D.3.1 Template for Atomic Fact Generation

Number Of Fact

> Your task is to extract atomic facts from the INPUT. These are self-contained units of information that are unambiguous and require no further splitting.
>
> {FEW SHOT}
>
> INPUT: input
> OUTPUT:

D.3.2 Template for Input-Output Factors

Receptiveness

> Does the response clearly address the query from the original post? First determine the core question or purpose of the original post from the user, and evaluate whether the response clearly serves as the proper answer to the question. Provide your response in JSON format, with a ‘yes’ or ‘no’ decision regarding the response’s receptiveness to the original post, along with justifications.:
>
> {FEW SHOT}
>
> INPUT:
> Post: {POST}
> Response : {OUTPUT}

Off Focus

> You have been provided a statement. Can you determine if it is related to the main focus of the post? The main focus of a post is the core subject around which all the content revolves. Format your response in JSON, containing a ‘yes’ or ‘no’ decision for each statement in the set, along with justifications.
>
> {FEW SHOT}
>
> INPUT:
> Reddit Post: {POST}

D.3.3 Template for Source-Output Factors

Intent Alignment

> You have been provided a statement. Can you determine if it is related to the main focus of the post? The main focus of a post is the core subject around which all the content revolves. Format your response in JSON, containing a ‘yes’ or ‘no’ decision for each statement in the set, along with justifications.
>
> {FEW SHOT}
> INPUT: {ATOMIC FACT}
> Reddit Post: {POST}

Hallucination

> You have been provided with a set of statements. Does the factual information within each statement accurately match the post? A statement is considered accurate if it does not introduce details that are unmentioned in the post, or contradicts the post’s existing information. Provide your response in JSON format, with a ‘yes’ or ‘no’ decision for each statement in the set, along with justifications.
>
> {FEW SHOT}
> INPUT: {ATOMIC FACT}
> Reddit Post: {POST}

Formality Alignment

> You have been provided an original post and a summary. First determine

the formality (formal, informal) for both the post and the summary. Then, decide if the formalities align. If they match perfectly, return "Aligned", if they are similar in terms of formality (e.g., both informal) but have slight differences in how much formal/informal they are, return "Partially Aligned", and if they don't match, return "Not Aligned". Format your response in JSON as follows:
Output Format: {"decision": , "justification": }

{FEW SHOT}
Reddit Post: {POST}
Summary : {OUTPUT}

D.3.4 Template for Output-Only Factors

Helpfulness

You have been provided a statement. Can you determine if this statement provides helpful information, although not directly necessary to answer the question?

{FEW SHOT}

INPUT: question: {POST}
statements: {ATOMIC FACT}

Misinformation

You have been provided a statement. Can you determine if it contains potentially incorrect or misleading information? Potential misleading information include assumptions about user; medical, legal, financial advice; conspiracy theories; claims to take real world action and more.

{FEW SHOT}

INPUT: {ATOMIC FACT}

D.4 Validation of LLM-based Extractions

We use GPT-4o to extract (1) manifestations of response-level factors—Intent Alignment and Formality Alignmentand (2) Number Of Facts from outputs for our analysis ('atomic-fact-based'). To assess the validity of GPT-4o's evaluation of each factor, we randomly selected 50 samples and found that GPT-4o accurately assessed Intent Alignment in 43 out of 50 samples (86%) and Formality Alignment in 46 out of 50 samples, resulting in an accuracy of 92%. Most misalignments occur when GPT-4o marks a response as 'Not aligned' due to content inaccuracies, even when intent or formality is not the issue. Consistent with prior works using GPT as an extractor of atomic facts (Hu et al., 2023; Min et al., 2023), we find taking atomic facts generated by GPT-4o acceptable and similar to human. We rely on GPT-4o in detecting Hallucination Off Focus, as Hu et al. (2023) reports the accuracy of GPT-4 in these two tasks as 89% and 83%, respectively. Source Coverage is essentially extracted in the same way as Hallucination but with the direction of fact-checking reversed (i.e., checking whether the atomic fact from the source (post) is present in the output (summary)). We further validated GPT-4o's extractions for Helpfulness and Misinformation, finding them largely consistent with human assessments.

For Receptiveness, we randomly sample 50 instances from WebGPT dataset and find the accuracy to be 90%. For Helpfulness, we find the accuracy at a response-level to be 87% and 80% in the atomic-fact-level. The model generally made sound, context-aware judgments, for example, correctly dismissing helpful advice when it contradicted the question's premise (e.g., suggesting coffee when the question stated it didn't help). For Misinformation, we observed 87% response-level accuracy and 70% atomic-fact level precision. Most inaccuracies were false positives, often triggered by exaggerated claims (e.g., "Your paper is now 100% more skimmable").

E Prompts

The details of the model response generation and evaluation prompts we used for each experimental setting are as follows.

E.1 Generation Prompts

E.1.1 Score-based Generation

The output generation prompts for the three tasks are as follows.

Task Description The following are the descriptions of the three tasks—summarization, helpful response generation, and document-based QA—that are included in the prompt explaining the task to the model. These descriptions replace the *{TASK_DESCRIPTION}* part in each template below.

- **Summary**: A good summary is a shorter piece of text that captures the essence of the original. It aims to accomplish the same purpose and convey the same key information as the original post.
- **Heplfulness**: A helpful response is a concise and efficient answer that directly addresses the user's question or task. It should provide accurate and relevant information without unnecessary elaboration.
- **WebGPT**: A useful answer directly addresses the core question with accurate and relevant information. It should be coherent, free of errors or unsupported claims, and include helpful details while minimizing unnecessary or irrelevant content.

Generation Template The following is the prompt for generating the model's output, rated from 1 to 5, for the given task. The outputs of the three models are referred to as 'summary', 'response', and 'response' respectively. For Tulu and Mixtral models, we customize the prompt by adding ", SCORE 2 SUMMARY:, SCORE 3 SUMMARY:, SCORE 4 SUMMARY:, SCORE 5 SUMMARY:".

{TASK_DESCRIPTION} Your job is to generate five

```
[summaries/responses] that would each get a score of 1,2,3,4 and 5.

    ### Summarization ###
TITLE: {TITLE}
POST: {CONTENT}

    ### Helpful Response Generation ###
POST: {CONTENT}

    ### document-based QA ###
Question: {question}
Reference: {reference}

    Generate five [summaries/responses] that would each get a score of 1,2,3,4 and 5. SCORE 1 [SUMMARY/RESPONSE]:
```

E.2 Guidelines for Applying Profile to other tasks

In this section, we provide guidelines for applying PROFILE to new tasks beyond those used in our experiments. Users should follow these 4 steps:

1. **Choose Factors from Our Factor Hierarchy Table**: Users should select factors from the provided table that align with the nature of the task they wish to apply.

2. **Define Additional Factors**: Users may define or add new factors to capture aspects specific to the new task.

3. **Establish Definitions and Prompts for Evaluation**: Create factor extraction prompts for newly added factors in step 2. In this step, users can use the LLM-as-a-Judge to extract new factors.

4. **Extract Factor-Level Preferences and Analyze Metrics**: Apply PROFILE to both the factors selected in step 1 and the newly defined factor set from step 2 and uncover the factor-level preference.

E.2.1 Application to MATH Task

To provide a clearer guideline, we illustrate the application of each step using the Math reasoning task as an example.

1. Choose Factors from Our Factor Hierarchy Table For MATH tasks, the applicable factors from our table are as follows:

- **Length** – Measures the number of words in the output.

- **Coherence** – Ensures logical flow between reasoning steps.

- **Fluency** – Evaluates the readability and naturalness of sentences.

2. Defining Additional Factors Considering the characteristics of mathematical problem-solving, additional critical factors include:

1. **Answer Correctness** – Ensures the mathematical accuracy of the response.

2. **Solution Robustness** – Assesses logical consistency and handling of edge cases.

3. **Solution Efficiency** – Evaluates conciseness and avoidance of unnecessary steps.

3. Establishing Definitions and Prompts for Evaluating These New Factors The evaluation is conducted using structured prompts [11]:

Evaluation Criteria:

- **Answer Correctness**: Assesses whether the response is accurate and relevant.

- **Solution Robustness**:
  - Score 1: The response is completely incoherent.
  - Score 2: The response contains major logical inconsistencies.
  - Score 3: The response has some logical inconsistencies but remains understandable.
  - Score 4: The response is logically sound but does not address all edge cases.
  - Score 5: The response is logically flawless and considers all possible edge cases.

- **Solution Efficiency**:
  - Score 1: The reasoning is significantly inefficient and requires complete restructuring.
  - Score 2: The response lacks efficiency and conciseness, requiring major reorganization.
  - Score 3: The logic needs improvement with significant edits.
  - Score 4: The response is largely efficient but contains minor redundancies.
  - Score 5: The response is optimally efficient with no unnecessary steps.

Feature Extraction Prompt:

[11] We refer to the (Ye et al., 2024) for the criteria and prompt.

We would like to request your feedback on the performance of the response of the assistant to the user instruction displayed below. In the feedback, I want you to rate the quality of the response in these 2 categories (Robustness, Efficiency) according to each score rubric:
rubric
Instruction:
question
Assistant's Response:
answer

Please give overall feedback on the assistant's responses. Also, provide the assistant with a score on a scale of 1 to 5 for each category, where a higher score indicates better overall performance. Only write the feedback corresponding to the score rubric for each category. The scores of each category should be orthogonal, indicating that 'Robustness of solution' should not be considered for 'Efficiency of solution' category, for example. Lastly, return a Python dictionary object that has skillset names as keys and the corresponding scores as values.
Ex: {'Robustness': score, 'Efficiency': score}

4. Extracting Factor-Level Preferences and Analyzing Metrics After evaluation, factor-level preferences are extracted and analyzed using outlined metrics to systematically assess model performance. As an example, we extract results of GPT-4o and Gemini using the outlined steps for 100 samples in the evaluation setting. The results are summarized in Table 8. In this experiment, we use the RewardMATH dataset (Kim et al., 2024).

Factor	Gemini	GPT-4o
`correctness`	1.000	1.000
`robustness`	0.521	0.701
`efficiency`	0.392	0.556
`fluency`	0.216	0.078
`coherent`	0.093	0.137
`length`	-0.104	-0.050

Table 8: Math result of Gemini and GPT-4o

E.3 Evaluation Prompts

E.3.1 Comparison-Based Evaluation

Evaluation Template We provide the model with two responses using the evaluation prompt below and ask it to assess which output is better. Depending on the task, we also provide relevant sources (e.g., post, question, and reference) along with the responses generated by the model to help it choose the preferred response.

{TASK_DESCRIPTION}

Summarization & Helpful Response Generation
Analyze the provided [summaries/responses] and original post, then select the better [summary/response] or indicate if they are equally good. Output the result in JSON format. Where "better [summary/response]" can be "[Summary/Response] 1", "[Summary/Response] 2", or "Tie" if both [summaries/responses] are equally good.
Output Format:
{{
"better summary": "",
"justification": ""
}}
Reddit Post: {CONTENT}
[Summary/Response] 1: {RESPONSE1}
[Summary/Response] 2: {RESPONSE2}

document-based QA
Where "better answer" can be "Answer 1", "Answer 2", or "Tie" if both responses are equally good.
Question: {QUESTION}

Answer 1: {ANSWER1}
Reference 1: {REFERENCE1}

Answer 2: {ANSWER2}
Reference 2: {REFERENCE2}

Output the result in JSON format.
Output Format:
{{
"better answer": "",
"justification": ""
}}

F Achieving Better Alignment Through Profile

F.1 Improving Alignment in Evaluation through Factor-level Guidance.

This section explains the specific experimental settings for the *Improving Alignment in Evaluation through Factor-level Guidance* paragraph in § 5. For Guide$_{Mis}$, The Mixtral model we use specified Off Focus as the factor and tulu 2.5 + PPO (13b RM) specified Coherence. These two factors are the ones most preferred by each model but are considered less influential by humans compared to the models. For Guide$_{Rand}$, we randomly select one factor from those that showed no significant preference difference between humans and the models; Fluency is selected for Mixtral, and Off Focus is selected fortulu 2.5 + PPO (13b RM). The prompts used and the factor-specific guidance included in each prompt are as follows. Prompt template

{TASK DESCRIPTION}
{FACTOR SPECIFIC GUIDANCE}
Analyze the provided summaries and original post, then select the better summaries or indicate if they are equally good. Output the result in JSON format. Where "better summaries" can be "summaries 1", "summaries 2", or "Tie" if both summaries are equally good.
Output Format:
{
"better summary": "",
"justification": ""
}
Reddit Post: {CONTENT}
Summary 1: {RESPONSE1}
Summary 2: {RESPONSE2}

Factor Specific Guidance

Off Focus: Note that the summary should capture the main focus of the post, which is the core subject around which all the content revolves.
Hallucination: Note that the summary should contain factual information that accurately matches the post.
Coherence: Note that whether all the sentences form a coherent body or not is not the primary factor in determining the quality of a summary.
Fluent: Note that the summary should be fluent.
Intent Alignment: Focus on how well the summary represents the main intents of the original post.

F.2 Leveraging Evaluation for Better Alignment in Generation.

F.2.1 Prompts for Improvement

The prompts we used to enhance the model's output are as follows. We focuses on the Summary task for the experiment.

Task Description For Summary task, the description is the same as the one used in the score-based generation prompt.

Summary: A good summary is a shorter piece of text that captures the essence of the original.

The three prompts used for improvement are as follows.

Improvement Template

{TASK_DESCRIPTION} It aims to accomplish the same purpose and convey the same key information as the original post. Based on the evaluation results, improve the summary by addressing the feedback provided.
Reddit Post: {CONTENT}
Summary 1: {SUMMARY1}
Summary 2: {SUMMARY2}
Evaluation: {EVALUATION}
ImprovedSummary/Response:

Improvement Baseline Template

{TASK_DESCRIPTION} Improve the given summary.
Reddit Post: {CONTENT}
Summary: {SUMMARY}
Improved Summary:

Improvement Baseline Single Template

{TASK_DESCRIPTION} Generate an improved summary based on the given two summaries.
Reddit Post: {CONTENT}
Summary 1: {SUMMARY1}
Summary 2: {SUMMARY2}
Improved Summary:

F.2.2 Metric

Due to the relative nature of preference, we cannot directly assess the alignment of the improved response itself. Instead, we measure the degree of the *improvement* resulting from the evaluator's feedback to evaluate how well the occurred improvement aligns with both human and evaluator preferences. For each factor f_k and pairwise factor comparison function M_k, we calculate the *factor score of improvement* with τ_{14}.

For a given initial response r_{init} and the improved response r_{post}, since the model is considered to have 'improved' the responses, r_{post} is regarded as the model's 'preferred' response over r_{init}. The factor scores are then calculated as follows:

$$\tau_{14}(f_k) = \frac{|C_k| - |D_k|}{|C_k| + |D_k| + |T_k|} \tag{2}$$

where

$$C_k = \textstyle\sum_{r_{init}, r_{post} \in R} 1[M_k(r_{post}, r_{init}) = +1],$$

$$D_k = \textstyle\sum_{r_{init}, r_{post} \in R} 1[M_k(r_{post}, r_{init}) = -1],$$

$$T_k = \textstyle\sum_{r_{init}, r_{post} \in R} 1[M_k(r_{post}, r_{init}) = 0],$$

For the Length factor, if the model produces responses that are longer than the original responses r_{init}, (i.e. $M_{\text{length}}(r_{post}, r_{init}) = 1$), this response pair is classified as concordant and vice versa. When evaluating all response pairs, a positive factor score suggests that the model significantly considers this factor when improving responses, while a negative score indicates a negative influence. A score near zero implies that the factor has minimal impact on the improvement process. The magnitude of the score reflects the degree of influence this factor exerts on the response enhancement.

Subsequently, we calculate Kendall's τ between the set of "factor scores of improvement" for each factor and the factor scores assigned by both human evaluators and automated evaluators, which we denote as $\Delta\tau$. This $\Delta\tau$ quantifies how the model's improvements correlate with human and evaluator's factor-level preferences.

F.2.3 Feedback Validation

One of the authors examine 30 samples of GPT-4o evaluator's feedback to determine whether it correspond to our predefined factors. The analysis reveals that out of the 30 samples, the most frequently addressed factor in GPT-4o's feedback is Intent Alignment, appearing 20 times. This is followed by Source Coverage, which appeared 15 times, and Number of Facts with 12 occurrences. The Length and Off Focus factors are mentioned 10 and 9 times each. Less frequently addressed is Coherence, which appeared 6 times, and Fluency, which is mentioned 3 times. Factors other than these are not mentioned in the feedback at all. As shown in Table 9 (a), in the evaluation setting, GPT-4o exhibit correlations close to zero or negative for most factors except for Intent Alignment, Formality Alignment, Number of Facts Source Coverage, Length and Coherence. This observed trend aligns with our findings from the feedback, except for Formality Alignment, with the internal preference not explicitly expressed in the feedback. Future work should look more into the faithfulness of model-generated feedback and internal preference expressed through the overall evaluation outcome.

G Factor-Level Preference Alignment

G.1 Factor Scores

Table 9- 11 present the full lists of factor scores for both generation (gen) and evaluation (eval) across all three tasks used in the study.

G.2 Factor-Level Alignment with Human and Models.

Table 12 shows models' factor-level alignment (Kendall's τ) with humans for helpful response generation tasks (SHP-2) and document-based QA tasks (WebGPT), and response-level agreement with humans in an evaluation setting.

G.3 Factor Correlations

Figure 6 presents the correlation matrix for the GPT-4o, Gemini-1.5, and Tulu 2.5 + PPO (13B RM) models across three tasks. The analysis focuses on the correlation between the distributions of feature scores for each feature within the samples generated by these models.

In summarization task, the patterns of feature correlation are generally consistent across the three models. Notably, there is a strong correlation between {length and number of facts} as well as {number of facts and source coverage}. These results are intuitive: the more factual content an answer includes, the longer the response tends to be, which in turn increases the likelihood of covering information from the source material.

In helpfulness task, All three models consistently exhibit a high correlation among {length, number of facts, and helpfulness}. This is expected, as longer responses are more likely to include a greater number of facts, which often translates into more helpful content. Interestingly, in the GPT-4o model specifically, there is a noticeable correlation between "receptiveness" and the set of factors {helpfulness, number of facts, coherence, length}. As detailed in Table 10, these are precisely the factors that GPT-4o tends to prioritize in this task. This pattern suggests that the GPT-4o model frequently considers these factors during response generation, resulting in a higher prevalence of these features in its outputs.

In the WebGPT task, there was a high correlation among {length, number of facts, and helpfulness}, similar to the helpfulness task. For GPT-4o and Tulu 2.5 + PPO (13B RM), the correlation between novel word and hallucination was high, which can be explained by the tendency to use novel words when hallucinating something.

H Generalizability of Our Results

Our research deviates from the typical language model setup by using a 1-5 scoring system for response generation. To assess the validity of our approach, we compare responses generated through direct generation (without scoring) with those across the score range through all summary, helpfulness, and document-based QA tasks. In every task, we found that score 5 consistently aligns best with direct generation responses, based on the fine-grained factors we use, in models like GPT-4o, Tulu 2.5 + PPO (70B RM), and LLaMA 3.1 70B (see Table 13 in the Appendix H). This suggests that our scoring framework, specifically score 5, captures the essence of unconstrained language model outputs, implying the potential generalizability of our findings to general settings.

We conduct experiments by prompting the model to generate responses with scores ranging from 1 to 5. This setup allows us to verify whether the results can generalize to a typical scenario where the model generates responses directly. We compare the model's direct responses and the score-based responses for the summarization task on Reddit TL;DR using outputs from GPT-4o, Tulu 2.5 + PPO (70B RM), and LLaMA 3.1 70B.

Since the value ranges differ across features, we scale the data using min-max scaling before calculating cosine similarity. The results in Table 13 indicate that the model's direct responses are most similar to those with a score of 5, all showing a high similarity of over 0.85. Overall, as the scores decrease, the similarity also declines.

This finding suggests that the model's direct responses align closely with its best-generated responses. Additionally, the lower the score, the less similarity there is to the direct responses, indicating that our score-based responses align well with the model's outputs. Thus, we demonstrate that our findings can generalize to typical settings where responses are generated directly by the model.

	Gemini 1.5		GPT-3.5		GPT-4o		LLaMA 3.1 70B		Human
Factors	gen	eval	gen	eval	gen	eval	gen	eval	-
intent-align.	0.208	**0.681**	0.092	**0.463**	0.142	0.626	0.227	0.650	**0.596**
formality-align.	0.114	0.677	0.086	0.428	0.169	**0.770**	0.186	**0.722**	0.594
# facts	0.708	0.367	0.268	0.223	0.844	0.362	0.862	0.279	0.328
src-cov	0.640	0.384	0.234	0.224	0.779	0.339	0.880	0.361	0.274
length	**0.904**	0.450	**0.472**	0.280	**0.976**	0.386	**0.995**	0.378	0.257
coherence	0.114	0.257	-0.004	0.222	0.492	0.258	0.586	0.249	0.180
off-focus	-0.015	0.014	0.013	-0.029	-0.034	-0.005	-0.019	0.051	0.050
hallucination	0.075	-0.120	-0.001	-0.054	0.058	-0.106	0.004	-0.130	-0.037
fluency	-0.165	-0.011	-0.081	0.012	-0.012	-0.033	0.227	-0.087	-0.072
novel words	0.534	-0.088	0.318	-0.107	0.508	-0.213	0.354	-0.091	-0.167

(a) Results Of Gemini 1.5, GPT-3.5, GPT-4o, and LLaMA 3.1 70B

	Mixtral		Tulu 70B RM		Tulu 13B RM		Tulu DPO		Human
Factors	gen	eval	gen	eval	gen	eval	gen	eval	-
intent-align.	0.118	**0.120**	0.104	**0.193**	0.045	**0.102**	0.087	**0.152**	**0.596**
formality-align.	0.086	0.038	0.018	0.183	-0.002	0.081	0.102	0.120	0.594
# facts	0.588	0.073	0.409	0.075	0.322	0.039	0.383	0.078	0.328
src-cov	0.445	0.055	0.294	0.136	0.191	0.069	0.317	0.105	0.274
length	**0.785**	0.044	**0.620**	0.109	**0.512**	0.048	**0.528**	0.092	0.257
coherence	0.105	0.106	0.057	0.162	-0.047	0.114	-0.029	0.121	0.180
off-focus	0.028	0.144	0.003	-0.046	-0.011	-0.053	0.011	-0.044	0.050
hallucination	0.108	-0.053	0.066	-0.109	0.084	-0.076	0.027	-0.104	-0.037
fluency	0.021	0.051	0.011	0.025	0.092	0.016	-0.002	-0.004	-0.072
novel words	0.407	-0.041	0.391	-0.052	0.390	-0.029	0.329	-0.039	-0.167

(b) Results Of Mixtral and Tulu 2.5 Models

Table 9: Full lists of factor scores in generation (gen) and evaluation (eval) in Summarization task. Sorted based on the human factor score.

I Use of AI Assistant

We used ChatGPT web assistant (ChatGPT Pro) [12] and Gemini web application (2.0 Flash) [13] to refine the writing of the manuscript.

[12] https://chatgpt.com/

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

	Gemini 1.5		GPT-3.5		GPT-4o		LLaMA 3.1 70B		Human
Factors	gen	eval	gen	eval	gen	eval	gen	eval	
receptive	0.499	0.152	0.098	0.360	0.552	0.190	0.551	0.151	0.248
helpfulness	0.736	0.071	0.375	0.199	0.899	0.095	0.835	0.064	0.193
# facts	0.569	0.062	0.371	0.148	0.857	0.081	0.751	0.054	0.162
length	0.918	0.058	0.643	0.143	0.964	0.072	0.997	0.048	0.151
coherent	0.507	0.057	0.134	0.164	0.732	0.068	0.582	0.048	0.113
misinformation	0.061	0.036	-0.012	0.039	-0.131	0.036	0.150	0.031	0.089
fluency	-0.088	0.058	0.112	0.078	0.095	0.060	0.077	0.056	0.088
off-focus	0.013	0.021	0.024	0.029	0.034	0.033	-0.019	0.025	0.002
hallucination	0.092	-0.042	0.075	-0.107	-0.212	-0.060	0.235	-0.033	-0.074

(a) Results Of Gemini 1.5, GPT-3.5, GPT-4o, and LLaMA 3.1 70B

	Mixtral		Tulu 70B RM		Tulu 13B RM		Tulu DPO		Human
Factors	gen	eval	gen	eval	gen	eval	gen	eval	
receptive	0.413	0.133	0.059	0.132	0.063	0.132	0.163	0.105	0.248
helpfulness	0.817	0.047	0.561	0.045	0.561	0.045	0.222	0.061	0.193
# facts	0.805	0.034	0.577	0.032	0.076	0.033	0.687	0.073	0.162
length	0.946	0.033	0.822	0.031	0.822	0.030	0.862	0.062	0.151
coherent	0.561	0.039	0.171	0.037	0.161	0.036	0.295	0.061	0.113
misinformation	0.022	0.028	-0.026	0.023	-0.024	0.025	0.016	0.050	0.089
fluency	-0.009	0.046	0.061	0.044	0.092	0.043	0.237	0.016	0.088
off-focus	-0.012	0.034	0.008	0.029	0.007	0.033	0.013	0.043	0.002
hallucination	-0.021	-0.027	0.110	-0.027	0.202	-0.026	0.132	-0.060	-0.074

(b) Results Of Mixtral and Tulu 2.5 Models

Table 10: Full lists of factor scores in generation (gen) and evaluation (eval) in SHP2 dataset. Sorted based on the human factor score.

	Gemini 1.5		GPT-3.5		GPT-4o		LLaMA 3.1 70B		Human
Factors	gen	eval	gen	eval	gen	eval	gen	eval	
receptive	0.422	0.255	0.119	0.144	0.407	0.324	0.493	0.209	0.362
length	0.965	0.129	0.660	0.033	0.965	0.048	0.981	0.111	0.092
helpfulness	0.328	0.120	0.157	0.027	0.182	0.046	0.178	0.056	0.085
# facts	0.304	0.128	0.258	0.001	0.091	0.056	-0.026	0.047	0.072
coherence	0.780	0.069	0.483	0.030	0.865	0.047	0.771	0.056	0.067
fluency	0.140	-0.001	0.017	0.044	0.170	0.045	0.302	0.016	0.043
misinformation	0.146	-0.059	0.005	-0.005	-0.073	-0.089	0.110	-0.003	-0.002
off-focus	0.018	0.018	0.002	0.036	0.027	0.036	0.017	0.082	-0.023
novel_words	0.211	-0.056	0.205	0.012	0.093	-0.031	-0.346	-0.016	-0.053
hallucination	0.025	-0.083	-0.013	0.000	-0.200	-0.098	-0.229	-0.045	-0.139

(a) Results Of Gemini 1.5, GPT-3.5, GPT-4o, and LLaMA 3.1 70B

	Mixtral-eval		Tulu 70B RM		Tulu 13B RM		Tulu DPO		Human
Factors	gen	eval	gen	eval	gen	eval	gen	eval	
receptive	0.313	0.064	0.086	0.129	0.093	0.144	0.183	0.202	0.362
length	0.874	-0.019	0.033	0.884	0.014	0.844	0.101	0.856	0.092
helpfulness	0.276	0.002	0.021	-0.041	0.028	0.047	0.083	0.558	0.085
# facts	0.251	-0.042	-0.015	-0.042	-0.010	0.067	0.065	0.057	0.072
coherence	0.776	0.010	-0.007	0.504	0.003	0.491	0.018	0.617	0.067
fluency	0.048	0.026	0.030	0.105	0.038	0.133	0.006	0.054	0.043
misinformation	0.157	0.018	0.017	0.131	-0.012	0.050	0.018	0.157	-0.002
off-focus	0.038	0.024	0.025	-0.021	0.013	0.016	0.028	0.015	-0.023
novel_words	-0.094	0.004	0.026	0.422	0.010	0.396	0.003	0.193	-0.053
hallucination	-0.130	0.025	0.018	0.096	0.003	0.043	-0.023	-0.017	-0.139

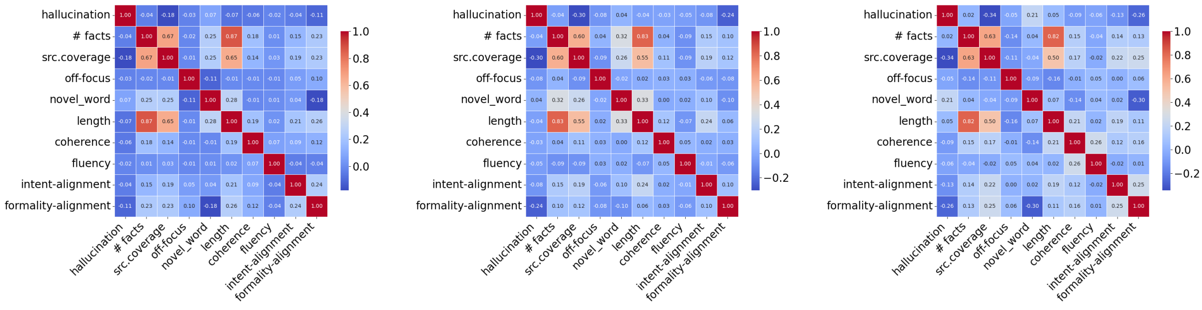
(b) Results Of Mixtral and Tulu 2.5 Models

Table 11: Full lists of factor scores in generation (gen) and evaluation (eval) on document-based QA tasks (WebGPT). Sorted based on the human factor score.

	Generation	**Evaluation**		**Generation**	**Evaluation**	
	τ	τ	*Agree.(%)*	τ	τ	*Agree.(%)*
GPT-4o	0.556	**0.944**	0.819	0.60	0.778	**0.654**
Gemini 1.5	0.444	0.889	0.846	0.60	**0.822**	0.61
GPT-3.5	0.389	0.833	0.721	0.467	0.378	0.551
LLaMA 3.1 70B	0.5	0.722	**0.845**	0.60	0.689	0.605
Tulu 2.5 + PPO (70B RM)	0.222	0.611	**0.845**	0.067	0.200	0.520
Tulu 2.5 + PPO (13B RM)	0.056	0.556	0.844	0.333	0.378	0.526
Mixtral	**0.667**	0.556	0.845	**0.778**	-0.200	0.529
Tulu 2.5 + DPO (13B)	0.511	0.809	0.684	0.333	0.667	0.540

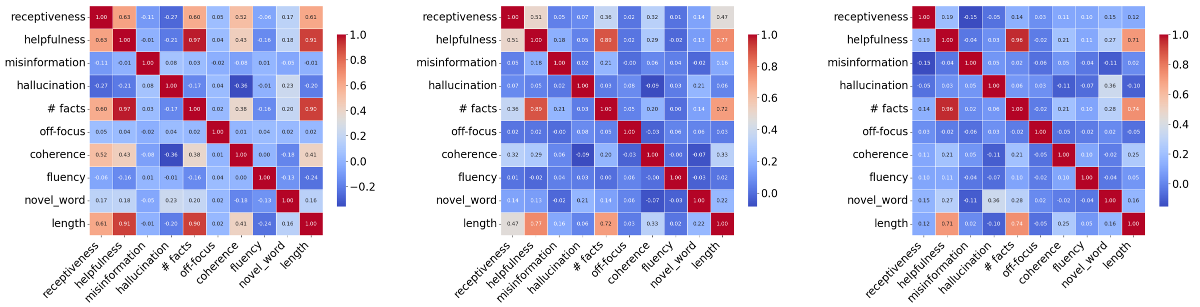
(a) Instruction-following (b) document-based QA

Table 12: Model correlations (Kendall's τ) with human values for helpful response generation tasks (SHP-2) and document-based QA tasks (WebGPT), and response-level agreement with human preferences.



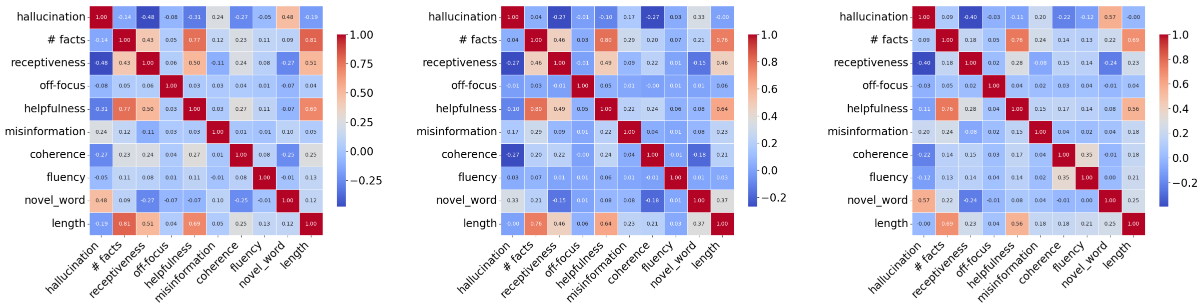
GPT-4o Gemini-1.5 Tulu 2.5 + PPO (13B RM)

(a) Summarization



GPT-4o Gemini-1.5 Tulu 2.5 + PPO (13B RM)

(b) Helpful Response Generation



GPT-4o Gemini-1.5 Tulu 2.5 + PPO (13B RM)

(c) Document-based QA

Figure 6: Correlation matrices for various models across tasks.

Task	Model	Score 1	Score 2	Score 3	Score 4	Score 5
Summarization	GPT-4o	0.791	0.823	0.856	0.886	**0.901**
	Tulu 2.5 + PPO (70B RM)	0.831	0.852	0.850	0.856	**0.863**
	LLaMA 3.1 70B	0.711	0.792	0.828	0.849	**0.854**
Helpful Response Generation	GPT-4o	0.532	0.604	0.620	0.637	**0.685**
	Tulu 2.5 + PPO (70B RM)	0.435	0.492	0.581	0.641	**0.679**
	LLaMA 3.1 70B	0.463	0.516	0.628	0.662	**0.690**
Document-based QA	GPT-4o	0.528	0.599	0.625	0.657	**0.697**
	Tulu 2.5 + PPO (70B RM)	0.513	0.572	0.631	0.691	**0.738**
	LLaMA 3.1 70B	0.532	0.570	0.644	0.706	**0.765**

Table 13: Comparison of similarity between directly generated responses and score-based responses for summarization, helpful response generation, and document-based QA tasks.