LLMs Become Logical Explainers through Persona-Conditioned Preference Optimization

Anonymous ACL submission

Abstract

Natural language explanations (NLEs) have been widely employed to convey relations underlying task outputs; however, existing approaches rely on superficial formulations, with limited use of theoretically grounded personas. In this study, we propose a multi-agent self-play framework in which a single model adopts multiple personas informed by diverse theories to generate NLE candidates. Relative preferences among these candidates are then induced through multi-factor scoring that jointly accounts for persona logical alignment, self-critique factuality from a critic agent, and expression diversity. These preferences are conveyed to a recomposer agent to synthesize a final NLE, which is subsequently used for self-preference learning, enabling the model to strategically learn from its own generations. Experiments suggest that our framework can improve logical alignment compared to single-persona approaches and reach coherent explanations, demonstrating that adaptive persona selection can be realized.

1 Introduction

Since the advent of LLMs, research has moved beyond producing task outputs toward generating natural language explanations (NLEs) that articulate the reasoning behind model predictions. Such NLEs are particularly valuable when task interpretation is inherently ambiguous or subjective, or when explanatory annotations require substantial human effort (He et al., 2024; Yao et al., 2023).

However, a principled understanding of how to guide models to generate reliable NLEs remains underdeveloped. In particular, LLM-based NLE generation is highly sensitive to the choice of reasoning strategy, and the vast design space makes identifying an optimal configuration challenging (Gemechu et al., 2025; Kunz and Kuhlmann, 2024). Moreover, prior work typically generated NLEs from a single perspective, limiting their ability to capture diverse interpretations in the absence of gold preferences (Agarwal et al., 2024).

In this paper, we introduce the **Self-Play Explainer (SPE)**, a multi-agent framework that decomposes a single LLM into multiple distinct reasoning agents, as shown in Figure 1. We define theory-grounded personas and analyze their preference relations in NLE generation, rather than treating personas as superficial role variations (Tseng et al., 2024). Our framework induces qualitatively distinct reasoning processes, such that personas effective for explaining simple factual relations may differ from those better suited for tasks requiring scientific background or contextual knowledge (Ahn et al., 2025). Although a universally optimal persona for NLE generation lacks explicit ground truth and remains difficult to define[1], we address this challenge through multi-factor preference scoring that aggregates relative preferences among persona-generated NLEs. This scoring jointly considers logical alignment between each NLE and the corresponding input-output pair (Scirè et al., 2024; Yang et al., 2024), critique-based penalties from a dedicated critic agent (Li et al., 2025), and expression diversity across generated NLEs.

Based on the induced preferences, we instruct a recomposer agent to synthesize a final NLE by integrating complementary elements from preferred candidates (Wang and Atanasova, 2025; Wang et al., 2025). This design is motivated by the observation that comprehensive NLEs arise from selectively integrating diverse perspectives rather than relying on a single persona. The resulting NLE and its aggregated scores are then used to construct preference pairs for supervised fine-tuning (SFT) and direct preference optimization (DPO) (Rafailov et al., 2023), yielding an explanation model aligned

[1]This challenge stems from the inherent subjectivity of human judgment (Chen et al., 2025), and the absence of guarantees that a single persona yields consistently aligned reasoning behaviors across different LLMs.

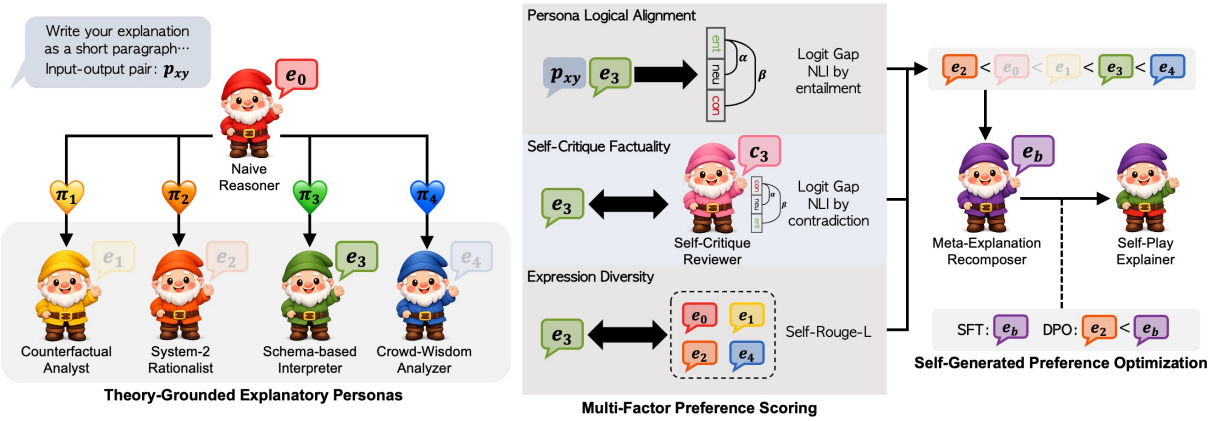

Figure 1: Overview of the proposed **Self-Play Explainer (SPE)** framework. For a given input-output pair, multiple theory-grounded personas generate candidate NLEs (§3.1), which are ranked using multi-factor preference scoring (§3.2). The top-ranked NLEs are then recomposed for preference optimization (§3.3).

with integrated persona reasoning.

Our experiments demonstrate that the proposed framework improves logical alignment over single-persona while preserving expression diversity within a controlled range. We observe gains in paragraph-level logical alignment alongside reduced sentence-level variability, resulting in more coherent NLEs. Further analyses indicate that the use of different personas yields distinct internal reasoning processes, and that the multi-factor scoring process is effective relative to the hard negative variant, highlighting the effectiveness of our framework for adaptive persona-driven NLE generation.

2 Related Work

2.1 The Role of Natural Language Explanations

NLEs have become a central component across various tasks, supporting interpretability and serving as auxiliary supervision. Early studies incorporated human-written explanations to justify predictions in tasks such as natural language inference (NLI) and question answering (QA) (Rajani et al., 2019; Camburu et al., 2018). Subsequent work has shifted toward automatic generation and explanation-aware training, positioning NLEs as a cross-task mechanism that supports reasoning, evaluation, and learning beyond prediction accuracy (Quan et al., 2025; Wang and Atanasova, 2025; Wadhwa et al., 2024).

Despite recent efforts, explanation-aware approaches have relied on fixed source NLEs that depend on a single reasoning trajectory, limiting their effectiveness across diverse contexts (Honda and Oka, 2025; He et al., 2024). In practice, NLEs often require different theoretical viewpoints depending on the context (Ahn et al., 2025). Motivated by this, we employ a persona-conditioned self-play paradigm within a single model to generate NLE candidates from multiple perspectives. This approach facilitates learning preference relations among them, thereby improving the logical consistency of NLEs through comparative optimization.

2.2 Persona Conditioning under a Perspectivist Framework

Early work employed personas to maintain conversational consistency in dialogue systems, inducing role-specific behavior through prompt design or fine-tuning (Zhang et al., 2018). Recent approaches have leveraged LLMs to elicit diverse perspectives; however, they described personas in terms of surface-level attributes, with fewer works defining them based on empirically grounded theoretical principles (Tseng et al., 2024).

We adopt a perspectivist view from cognitive and social sciences, which characterizes NLEs as inherently dependent on theoretical viewpoints rather than as absolute accounts. Importantly, in our study, perspectivism does not imply arbitrariness but instead recognizes that different viewpoints can highlight compatible aspects of the reasoning process within a single model (Lombrozo, 2006).

2.3 Recent Advances in Multi-Agent Reasoning

Research on multi-agent reasoning has emerged to address the limitations of single-pass inference in complex tasks. Early studies focused on agent decomposition, assigning specialized roles such

as planning or verification to enhance reasoning quality (Ma et al., 2025; Hong et al., 2023). Subsequent approaches leveraged agent interactions to explore alternative reasoning paths and enable mutual refinement (Karthikeyan et al., 2025).

Unlike prior studies that primarily employ agents as functional modules for problem solving or verification, we reconceptualize multi-agent reasoning as a mechanism for inducing NLE preferences within a single model. Persona-conditioned agents are utilized to generate, compare, and synthesize NLE candidates (Li et al., 2025; Wang et al., 2025), thereby converting relative preferences into learning signals that facilitate adaptive NLE generation.

3 Self-Play Explainer Framework

3.1 Theory-Grounded Persona Definition

We design five personas with distinct reasoning policies to generate NLEs. In contrast to prior approaches that rely on superficial patterns (Honda and Oka, 2025; He et al., 2024), we hypothesize that strategically designed personas enable a single model to elicit diverse behavior. In this setting, multiple reasoning paths are expected to reveal logical evidence that a single NLE may overlook, while their comparison facilitates evaluation of the underlying model structure.

Each persona operates as follows: **(1) Naive Reasoner**, relying solely on the model's knowledge without external theoretical grounding; **(2) System-2 Rationalist**, explicitly expressing chain-of-thought (CoT) reasoning aligned with Kahneman's dual-process theory (Kahneman, 2011); **(3) Counterfactual Analyst**, applying 'what-if' contrasts inspired by Pearl's causal framework (Pearl, 2009); **(4) Schema-based Interpreter**, providing structured interpretations grounded in Bartlett's schema theory (Bartlett, 1995); and **(5) Crowd-Wisdom Synthesizer**, integrating multiple perspectives following Surowiecki's principle of collective intelligence (Surowiecki, 2005). More descriptions of each persona are provided in Appendix A.1, and their prompts are in Appendix D.

We select personas that are broadly applicable across domains and design our framework to adaptively select appropriate personas for each input. We treat persona choice as a controlled variable, leaving broader persona exploration to future work.

3.2 Multi-Factor Preference Scoring

We establish systematic criteria for identifying preferable personas, recognizing that the optimal persona choice depends on the reasoning demands. Since a single, universally applicable ground truth NLE preference is difficult to define, we focus instead on distinguishing *relatively preferred* NLEs among persona candidates.

We define p_{xy} as the task-specific formulation text for the x-y pair, and the NLE generated by the k-th persona as e_k. Here, x and y denote the task input and output, π_k denotes the persona-specific reasoning policy, and the model is shared across personas. We denote by E the set of NLEs generated by the personas for the same p_{xy}.

$$e_k \sim \mathrm{LLM}(e \mid \mathsf{p}_{xy}, \pi_k), \quad (1)$$

$$\mathrm{E} = \{e_k \mid k = 1, ..., K\}, \quad (2)$$

3.2.1 Persona Logical Alignment

Recent studies have adopted NLI-based scores to evaluate the quality of generated texts (Scirè et al., 2024; Yang et al., 2024). However, as such approaches primarily rely on entailment signals, they often fail to capture subtle neutrality and contradiction patterns among persona-generated NLEs. To address this, we introduce a logit gap measure that incorporates the remaining logits. We define a ***flip*** when two NLEs have similar entailment (ENT) logits but their preference is reversed once neutral (NEU) and contradictory (CON) logits are considered; when this occurs under only a marginal logit difference, we refer to it as a ***near-tie flip***.

First, we define $\ell_{xy,k}$ as the NLI logit for a given class between p_{xy} and e_k.

$$\ell_{xy,k}^{\{\text{ENT},\text{NEU},\text{CON}\}} = \mathrm{NLI}^{\{\text{ENT},\text{NEU},\text{CON}\}}(\mathsf{p}_{xy}, e_k), \quad (3)$$

In measuring the logit gap, we adaptively calibrate the hyperparameters (α, β) to the given training data and model rather than treating them as fixed constants. We then select the values that capture meaningful *near-tie flips* through grid search. The NLI model, formal definitions of *(near-tie) flips*, grid search procedure, and selected (α, β) pairs are provided in Appendix A.2.

Based on this, we define the entailment logit gap between p_{xy} and e_k as follows:

$$gap_{xy,k}^{(\text{ENT})} = \ell_{xy,k}^{(\text{ENT})} - \alpha \cdot \ell_{xy,k}^{(\text{NEU})} - \beta \cdot \ell_{xy,k}^{(\text{CON})}, \quad (4)$$

We apply softmax over the logit gap across different NLEs to model their relative preference, shifting

the probability distribution from class to candidate-wise competition. To avoid distortion from absolute logit scale differences, we apply z-score normalization to the logit gaps before softmax.

$$S_{\mathsf{PLA}}(\mathsf{p}_{xy}, e_k, \mathrm{E}) = \underset{k}{\mathrm{softmax}}(\widetilde{gap}^{(\mathrm{ENT})}_{xy,k}) \quad (5)$$

This enables a more fine-grained assessment of each persona's logical alignment, even when entailment scores are similar, by flexibly incorporating the remaining logits informed by the data and model characteristics used in NLE generation.

3.2.2 Self-Critique Factuality

Recent approaches to evaluating the factuality of LLM-generated content have prompted models to produce self-assessment scores or leverage internal feature signals (Zhang et al., 2024b; Parcalabescu and Frank, 2024). However, these methods often suffer from limited interpretability. To mitigate this, we introduce a **`Self-Critique Reviewer`** agent that enables claim-level critiques of its own generated NLEs (Li et al., 2025) and incorporates contradiction-focused penalties to directly capture inconsistencies.

We assign two subtasks to this agent: to classify the factual relationship between p_{xy} and e_k using one of three scales, and to generate a claim-level critique consistent with the selected scale (Chu et al., 2025). The claim-level critique generated by the k-th persona, denoted c_k, is defined as follows, where π_{crit} denotes the self-critique reasoning policy.

$$c_k \sim \mathrm{LLM}(c \mid \mathsf{p}_{xy}, e_k, \pi_{\text{crit}}), \quad (6)$$
$$\mathrm{C} = \{c_k \mid k = 1, ..., K\}, \quad (7)$$

We segment e_k and c_k into sentences, denoted as $e_k = \{s_1, ..., s_N\}$ and $c_k = \{t_1, ..., t_M\}$. To identify counter-argument pairs, we apply the logit gap measure under the *near-tie flip* condition defined as contradiction. The hyperparameters (α, β) are adaptively determined by the training data and model. Based on this, we define the contradiction logit gap between s_n and t_m, where $n \in [1, N]$ and $m \in [1, M]$, as follows:

$$gap^{(\mathrm{CON})}_{n,m} = \ell^{(\mathrm{CON})}_{n,m} - \alpha \cdot \ell^{(\mathrm{NEU})}_{n,m} - \beta \cdot \ell^{(\mathrm{ENT})}_{n,m}, \quad (8)$$

We focus on the distribution tail capturing the strongest counterarguments, selecting cases with the most pointed critiques. Among the $N \times M$ sentence pairs, we define $\mathcal{T}_q$ as the subset whose contradiction logit gaps fall within the top $q\%$, and compute the mean over this set.

$$\mathcal{C}(e_k, c_k) = \frac{1}{|\mathcal{T}_q|} \sum_{n,m \in \mathcal{T}_q} gap^{(\mathrm{CON})}_{n,m}, \quad (9)$$

We apply softmax over the logit mean across different NLEs to model their relative preference, including z-score normalization before softmax.

$$S_{\mathsf{SCF}}(e_k, c_k, \mathrm{E}, \mathrm{C}) = \underset{k}{\mathrm{softmax}}(\widetilde{\mathcal{C}}(e_k, c_k)), \quad (10)$$

This enables a more controlled evaluation by requiring the model to generate a self-critique proxy from a strictly critical perspective, with contradiction-focused penalties imposed.

3.2.3 Expression Diversity

We expect different personas to exhibit independent reasoning styles, even when explaining the same x-y relation. Thus, we extend similarity-based diversity assessment (Zhang et al., 2025; Zhu et al., 2018) to the paragraph-level and interpret higher cross-persona similarity as relatively low expression diversity. In this setting, we employ Rouge-L, which captures shared narrative structure through the longest common subsequence.

$$S_{\mathsf{ED}}(e_k, \mathrm{E}) = \frac{1}{K-1} \sum_{k' \neq k} \text{Rouge-L}(e_k, e_{k'}), \quad (11)$$

However, length imbalance among NLEs may distort similarity evaluation. To mitigate this, we apply a length-based weighting that down-weights pairs with large discrepancies, using a Laplacian kernel to convert length differences into an attenuation factor that decreases a pair's contribution as the length gap increases (Rahimi and Recht, 2007)[2].

$$w_{kk'} = \exp(-\gamma \big| \, |e_k| - |e_{k'}| \, \big|), \quad (12)$$

We incorporate $w_{kk'}$ into the original scheme, enabling the assessment of independent reasoning styles by evaluating the diversity of multiple NLEs while accounting for length bias.

$$S_{\mathsf{ED}}(e_k, \mathrm{E}) = \frac{\sum_{k' \neq k} w_{kk'} \cdot \text{Rouge-L}(e_k, e_{k'})}{\sum_{k' \neq k} w_{kk'}}, \quad (13)$$

[2]We set $\gamma = 0.04$, so that a one-sentence length difference halves the weight of a pair of NLEs. Since all personas were instructed to follow the same length constraint, such deviations may also indicate inconsistency in instruction following.

Dataset	HotpotQA		NQopen		ANLI						eSNLI	
					R1		R2		R3			
Metric	$P^{(\text{ENT})}$ ↑	S_{ED} ↓	$P^{(\text{ENT})}$ ↑	S_{ED} ↓	$P^{(\text{ENT})}$ ↑	S_{ED} ↓	$P^{(\text{ENT})}$ ↑	S_{ED} ↓	$P^{(\text{ENT})}$ ↑	S_{ED} ↓	$P^{(\text{ENT})}$ ↑	S_{ED} ↓
`System-2 Rationalist`	48.50	35.48	41.54	35.36	47.02	36.42	44.88	36.41	43.79	35.10	36.75	36.02
`Counterfactual Analyst`	57.56	33.46	54.22	34.72	45.77	32.03	44.48	32.19	43.69	31.35	36.34	32.50
`Schema-based Interpreter`	30.00	35.93	31.24	38.20	30.32	36.89	29.84	36.77	32.53	36.11	25.25	37.03
`Crowd-Wisdom Synthesizer`	69.84	30.09	63.89	30.84	44.64	31.22	45.50	31.25	**53.15**	30.25	48.27	29.51
Best-of-Personas	91.04	27.54	88.92	28.41	81.29	28.07	81.12	28.06	82.41	27.10	77.47	27.41
Self-Play Explainer (SFT)	50.02	32.42	54.04	35.82	39.99	20.00	40.46	39.28	39.37	38.14	39.33	29.81
Self-Play Explainer (DPO)	**75.32**	40.19	**75.94**	41.41	**54.01**	41.14	**53.08**	41.70	52.32	43.54	**62.29**	44.13

Table 1: Performance comparison of the **Self-Play Explainer (`SPE`)** framework under the persona-conditioned setting, using Llama-3.2-3B(-Instruct) as the base model. $P^{(\text{ENT})}$ denotes entailment probability, where higher values indicate better logical consistency (We bold the best score for each task excluding Best-of-Personas), while lower S_{ED} indicates higher expression diversity. We report the average scores computed over three sampling runs for diversity scores.

3.3 Self-Generated Preference Optimization

We compute a harmonic mean to discourage candidates that perform well on only a single score and instead prefer `NLE`s with balanced quality. Since higher values are preferred for S_{PLA}, whereas lower values are preferred for S_{SCF} and S_{ED}, this relationship is reflected in the formulation. Although score weights can be adjusted for specific applications, we use equal weighting in this study.

$$S_{\texttt{Final}} = \frac{3}{\frac{1}{S_{\text{PLA}}} + \frac{1}{1-S_{\text{SCF}}} + \frac{1}{1-S_{\text{ED}}}}, \qquad (14)$$

Rather than relying on the first-ranked result, we assume that aggregating diverse perspectives yields a more comprehensive `NLE`. To this end, we introduce a **`Meta-Explanation Recomposer`** agent that identifies commonly supported evidences across the top-2 `NLE`s and generates a single final `NLE` (Wang and Atanasova, 2025; Wang et al., 2025). To incorporate the final preferences, we define the `NLE` produced by the recomposer agent as e_{win} and the one with the lowest $S_{\texttt{Final}}$ as e_{lose}. We then perform SFT to generate e_{win} for each x-y pair, and apply DPO with $(e_{\text{win}}, e_{\text{lose}})$ pairs to encode the preference signal. The prompts for all agents employed in our framework are presented in Appendix D.

4 Experimental Results

4.1 Tasks

We select extractive QA and NLI because both tasks require multi-step evidence-based reasoning or semantic relations. We use HotpotQA, which involves multi-document reasoning (Yang et al., 2018), and NQopen, a short-answer variant of Natural Questions reflecting user queries (Lee et al., 2019). We also use ANLI across its R1-3 splits with increasing reasoning difficulty (Nie et al., 2020) and eSNLI, which provides simple annotations as part of supervision (Camburu et al., 2018).

4.2 Evaluation Metrics and Setup

We measure the logical alignment between each task input-output pair and its generated `NLE`, treating them as a premise-hypothesis pair to compute the entailment probability from a pre-trained NLI model[7]. This design follows recent approaches that leverage NLI scores as a proxy for faithfulness and logical coherence (Zhang et al., 2024a). We then employ the S_{ED} score to evaluate the expression diversity under a fixed sampling. Further details on dataset curation and training configuration are provided in Appendix B.

4.3 Main Results

We report the performance comparison in Table 1, using Llama-3.2-3B(-Instruct) as the base model. `Crowd-Wisdom Synthesizer` attains the highest logical alignment, whereas `Schema-based Interpreter` performs the worst, with the largest logical alignment gap of +39.84 observed on HotpotQA. The remaining personas generally show intermediate performance; while their results are comparable on NLI tasks, `Counterfactual Analyst` demonstrates a clear advantage on QA tasks over `System-2 Rationalist`. This indicates that the effectiveness of personas for `NLE` generation depends on task-specific characteristics.

Under the Best-of-Personas, all personas are sampled and the highest-scoring instance is selected (Rafailov et al., 2023), yielding the highest scores. However, this approach is computationally impractical, as it requires K sampling completions and corresponding metric evaluations per instance. In this case, **`SPE`** (DPO) achieves results closest to

Dataset	Hotpot QA	NQ open	ANLI			eSNLI
			R1	R2	R3	
SPE (SFT)	50.02	54.04	39.99	40.46	39.37	39.33
SPE (DPO)	**75.32**	**75.94**	**54.01**	**53.08**	**52.32**	**62.29**
SPE (DPO) w/o SFT	42.16	35.62	35.24	34.83	40.19	37.25

Table 2: Logical alignment scores of the **SPE** framework in case of the SFT and DPO training, using Llama-3.2-3B(-Instruct) as the base model. We bold the best score for each task.

Dataset	Hotpot QA	NQ open	ANLI			eSNLI
			R1	R2	R3	
CoT	67.03	55.43	53.25	52.98	**55.93**	48.67
self-refine (`System-2`)	62.04	58.24	50.06	**53.26**	53.49	51.22
self-refine (`Counter`)	61.82	62.42	46.13	45.77	47.12	45.14
self-refine (`Schema`)	44.54	44.18	39.95	39.50	42.42	38.67
self-refine (`Crowd`)	64.00	61.81	45.24	45.96	50.12	46.36
SPE (DPO)	**75.32**	**75.94**	**54.01**	53.08	52.32	**62.29**

Table 3: Logical alignment scores of the **SPE** framework with the supplementary baselines, using Llama-3.2-3B(-Instruct) as the base model. We bold the best score for each task.

Best-of-Personas among the compared methods, indicating the effectiveness of preference learning across personas[3]. Regarding expression diversity, each persona exhibits comparable scores, with `Crowd-Wisdom Synthesizer` showing relatively lower values, likely due to its expert-role simulation encouraging a more narrative reasoning style. Compared to **SPE** (SFT), the score increases after applying DPO, reflecting an alignment-diversity trade-off arising from shifts in the model's token distribution toward persona-specific behaviors.

The results show that **SPE** (DPO) improves logical alignment across all tasks, outperforming most individual personas. Notably, it achieves gains of +25.30 on the HotpotQA task and +12.62 on the ANLI-R2 compared to **SPE** (SFT). To further examine the performance degradation observed under SFT-only training in our framework, we additionally perform direct DPO without prior SFT and report the results in Table 2. Across all tasks, DPO achieves its best performance when preceded by SFT, suggesting that SFT primarily shapes persona-conditioned representations rather than directly improving performance, consistent with prior observations (Rafailov et al., 2023). Although **SPE** (SFT) alone yields modest results, it effectively aligns

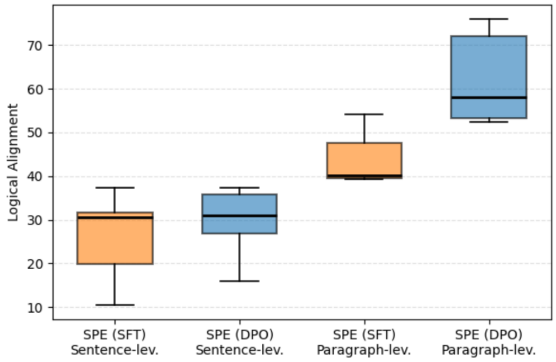

Figure 2: Boxplot distributions of sentence- (left) and paragraph-level (right) logical alignment scores for SFT (orange) and DPO (blue) models. The results are aggregated across all tasks in our experiments[4].

persona-informed `NLE`s within the model's behavioral space, thereby establishing a foundation that facilitates subsequent **SPE** (DPO) training.

4.4 Supplementary Baseline Results

We also include CoT and self-refine without persona conditioning as supplementary baselines (Madaan et al., 2023; Wei et al., 2022), and report the results in Table 3. Although these methods achieve higher scores than individual personas, the proposed framework attains the highest scores on four out of six tasks. While the margin remains modest on the remaining tasks, competing methods exhibit inconsistent improvements across other tasks. These results suggest that the observed gains arise not merely from internal reasoning or iterative refinement, but from more effectively integrating diverse perspectives in `NLE` generation.

5 Discussion

5.1 Granularity Effects in Logical Alignment

Our main results measure logical alignment of `NLE`s at the paragraph-level. To analyze alignment at finer granularity, we additionally compute sentence-level alignment by decomposing each `NLE` into sentences and evaluating them against the input-output pair. This enables a more precise characterization of whether observed gains stem from individual sentence-level accuracy or from coherent organization at the paragraph-level. The results are presented in Figure 2.

While **SPE** {SFT, DPO} exhibit comparable sentence-level alignment, with average scores of

[3]Persona distributions favored by Best-of-Personas vary across tasks, and are further discussed in Appendix C.3.

[4]Detailed sentence-level scores for each task are reported in Appendix C.4.

GEMBA score	Naive	System-2	Counter	Schema	Crowd
Naive	2.57 ± 0.98	2.95 ± 1.05	**3.67** ± 0.91	3.23 ± 0.90	3.21 ± 0.92
System-2	2.95 ± 1.05	2.86 ± 1.09	**3.74** ± 0.94	3.34 ± 0.90	3.34 ± 0.98
Counter	3.67 ± 0.91	3.74 ± 0.94	3.25 ± 0.96	**3.81** ± 0.86	3.56 ± 0.94
Schema	3.23 ± 0.90	3.34 ± 0.90	**3.81** ± 0.86	2.51 ± 0.99	3.31 ± 0.93
Crowd	3.21 ± 0.92	3.34 ± 0.98	**3.56** ± 0.94	3.31 ± 0.93	2.90 ± 0.90

Table 4: GEMBA scores evaluating the degree of reasoning differences. Higher scores indicate more divergent reasoning; the highest score in each comparison pair is highlighted in bold. We report the grey-box values indicating comparisons between different sampling runs under the same persona.

Dataset	Hotpot QA	NQ open	ANLI			eSNLI
			R1	R2	R3	
System-2	37.31	35.67	27.05	28.11	32.71	24.86
Counter	29.89	27.51	22.41	22.74	23.70	11.07
Schema	19.25	42.37	25.96	25.81	28.73	22.87
Crowd	**51.78**	37.05	48.19	47.51	55.45	16.08
CoT	33.25	18.99	32.68	33.00	42.19	40.80
self-refine (System-2)	35.97	26.19	27.78	27.92	37.76	34.47
self-refine (Counter)	33.21	27.25	27.00	26.30	34.37	31.93
self-refine (Schema)	34.70	28.22	24.73	24.82	33.22	28.85
self-refine (Crowd)	33.46	20.91	23.27	23.91	33.28	36.93
SPE (SFT)	48.73	47.11	**54.37**	**53.15**	57.43	41.55
SPE (DPO)	49.54	**49.93**	53.97	52.94	**57.50**	**42.50**

Table 5: Logical alignment scores of the **SPE** framework, using Qwen2.5-7B(-Instruct) as the base model. We bold the best score for each task.

26.18 and 29.68 respectively, **SPE** (DPO) achieves substantially higher alignment at the paragraph-level, achieving an average score of 62.16 compared to 43.86 for **SPE** (SFT). This pattern suggests that **SPE** (DPO) improves global coherence and cross-sentence consistency. Moreover, it exhibits reduced variability in sentence-level scores, indicating more stable behavior and demonstrating its effectiveness in aligning NLEs.

5.2 Reasoning Differences Analysis

To further verify whether different personas lead to changes in the reasoning process, we employ a GEMBA-style evaluation using GPT-4.1-mini to assess the degree of reasoning difference between two NLEs (Kocmi and Federmann, 2023). We compared outputs with the same and different personas from multiple sampling runs, and report the results in Table 4.

Scores obtained from different personas show consistently larger values than those from the same persona, suggesting that different persona settings induce more diverse reasoning behavior. To confirm that this difference is not due to chance, we additionally perform a paired Wilcoxon signed-rank test to verify statistically significant differences across personas. In this case, the null hypothesis (H_0) states that the median difference between GEMBA scores from the different (A) and same (B) personas is less than zero.

$$H_0 : \text{median}(A - B) \leq 0, \tag{15}$$

For all ten pairwise comparisons between different personas, the score differences were statistically significant, resulting in rejection of H_0 with $p < 0.01$. In particular, Naive Reasoner and Counterfactual Analyst showed the largest mean score difference of 1.09, providing evidence of substantially different reasoning trajectories.

5.3 Model Coverage Analysis

We further trained different models using the proposed framework, and report the results in Table 5. In Qwen2.5-7B, unlike the preceding model, individual personas did not exhibit consistently high or low logical alignment scores across tasks. For instance, whereas Crowd-Wisdom Synthesizer achieved consistently strong performance across all tasks in Llama3.2-3B, this tendency was not observed in this model, where performance disparities became more pronounced depending on the task-specific persona pairings. The supplementary baselines tended to underperform compared to individual personas, suggesting that the effectiveness of NLE generation strategies may vary depending on the underlying model architecture[5].

Consistent with the earlier observations, smaller models showed lower logical alignment scores for **SPE** (SFT) alone than for individual personas[5]. In contrast, **SPE** (SFT) outperformed most baselines. This suggests that SFT may contribute to performance improvements in larger models even when primarily serving to align the model's internal representations prior to DPO optimization. In contrast, the performance gains obtained after applying **SPE** (DPO) were limited to three tasks, suggesting that effective logically grounded NLE generation may require the careful selection of model architecture.

[5]These patterns were also observed for Qwen2.5-3B.

Dataset	HotpotQA		NQopen		ANLI						eSNLI	
					R1		R2		R3			
Metric	$P^{(\text{ENT})} \uparrow$	$S_{\text{ED}} \downarrow$	$P^{(\text{ENT})} \uparrow$	$S_{\text{ED}} \downarrow$	$P^{(\text{ENT})} \uparrow$	$S_{\text{ED}} \downarrow$	$P^{(\text{ENT})} \uparrow$	$S_{\text{ED}} \downarrow$	$P^{(\text{ENT})} \uparrow$	$S_{\text{ED}} \downarrow$	$P^{(\text{ENT})} \uparrow$	$S_{\text{ED}} \downarrow$
`Schema-based Interpreter`	30.00	35.93	31.24	38.20	30.32	36.89	29.84	36.77	32.53	36.11	25.25	37.03
`Crowd-Wisdom Synthesizer`	69.84	30.09	63.89	30.84	44.64	31.22	45.50	31.25	53.15	30.25	48.27	29.51
Self-Play Explainer (SFT)	50.02	32.42	54.04	35.82	39.99	20.00	40.46	39.28	39.37	38.14	39.33	29.81
Self-Play Explainer (DPO)	75.32	40.19	75.94	41.41	54.01	41.14	53.08	41.70	52.32	43.54	62.29	44.13
Self-Play Explainer (SFT) w Hard Negative	50.43	33.21	46.49	33.82	37.32	35.08	37.38	35.01	41.09	34.03	28.25	28.27
Self-Play Explainer (DPO) w Hard Negative	32.81	44.18	27.76	39.23	27.81	33.00	36.05	36.01	41.79	33.64	21.79	28.14

Table 6: Performance comparison of the **Self-Play Explainer (SPE)** framework and its hard negative variant, using Llama-3.2-3B(-Instruct) as the base model. Here, the single-persona baselines correspond to the two personas with the highest and lowest logical alignment across all tasks in the original results in Table 1.

The results for Qwen2.5-3B, along with the hyperparameter tuning for training the larger 7B model, are presented in Appendix C.5.

5.4 Ablation Study

To evaluate the functional attribution of the proposed framework, we conduct an ablation study with hard negative variants. In this setting, e_{win} is defined as the persona ranked fourth according to the multi-factor scoring, while e_{lose} is selected using the same procedure as in the prior setting. The results are reported in Table 6.

Under the standard framework, logical alignment improves from **SPE** (SFT) to (DPO) training, whereas the hard negative variant exhibits degraded performance. As SFT progresses with hard negative variants, scores fall below those of the strongest single persona, `Crowd-Wisdom Synthesizer`. This degradation is further amplified after DPO training and, on NQopen, logical alignment drops by up to -3.48 relative to the weakest single persona, `Schema-based Interpreter`. Although diversity scores show improvements across tasks, they are accompanied by a substantial decline in logical alignment.

These results indicate that aggressively constructed preference signals can impair `NLE` optimization, resulting in excessive degradation in logical alignment. We demonstrate that strategically leveraging the relations induced by multi-factor scoring enables effective **SPE** (DPO) training that improves the logical consistency of `NLE`s[6].

6 Conclusion

We propose a multi-agent self-play framework based on a single model, where `NLE`s are generated from theory-grounded perspectives for the same input-output pair. These `NLE`s are evaluated using a multi-factor scoring that measures persona logical alignment, self-critique factuality with the reviewer agent, and expression diversity to derive relative preferences. The recomposer agent then synthesizes a final `NLE` and enables self-optimization through DPO training. Experimental results demonstrate that our framework can improve logical alignment across tasks compared to single-persona baselines. While no single persona baseline consistently dominates across tasks, our framework learns to leverage suitable personas during training.

Limitations

Persona generalization Our study assumes the utility of persona-conditioned `NLE`s, without explicitly modeling when personas are required or dynamically assigning them to specific contexts. Instead, by jointly evaluating multiple personas, our approach enables self-optimization of `NLE` generation and captures adaptive persona-selection behavior. Further exploration is required to assess whether alternative personas may be more effective in other domains.

Metric dependency In the absence of ground truth, our approach relies on multi-factor scoring to infer relative preferences rather than implying an optimal metric aligned with human annotations, primarily relying on NLI-based scores. For more rigorous evaluation, future work is required to develop dedicated metrics that explicitly reflect persona-conditioned preferences. We also report GPT-4 win rates in Appendix C.6, along with an analysis of its limitations for persona evaluation.

Computational cost Although our framework introduces training-time overhead from repeated multi-agent inference, we observe stable learning even with the relatively small model, and expect

[6] We conducted a human evaluation as a sanity check for human-perceived groundedness to assess whether the `NLE`s are well-grounded from a human perspective; the results suggest this tendency, with details reported in Appendix C.1.

similar trends in more diverse models.

Ethics Statement

The personas used in our study are intended as modeling tools and do not represent real individuals or social groups and are not tied to demographic attributes. We used an LLM with model outputs constrained by explicitly defined system prompts, and the model was not deployed in ways that introduce safety risks. All datasets used in this work are publicly available and do not contain personal or sensitive user information.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

Yongsu Ahn, Yu-Ru Lin, Malihe Alikhani, and Eunjeong Cheon. 2025. Human-centered explanation does not fit all: The interplay of sociotechnical, cognitive, and individual factors in the effect ai explanations in algorithmic decision-making. *arXiv preprint arXiv:2502.12354*.

Frederic Charles Bartlett. 1995. *Remembering: A study in experimental and social psychology*. Cambridge university press.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10777–10802, Vienna, Austria. Association for Computational Linguistics.

Zheng Chu, Huiming Fan, Jingchang Chen, Qianyu Wang, Mingda Yang, Jiafeng Liang, Zhongjie Wang, Hao Li, Guo Tang, Ming Liu, and Bing Qin. 2025. Self-critique guided iterative reasoning for multi-hop question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2415–2438, Vienna, Austria. Association for Computational Linguistics.

Debela Gemechu, Ramon Ruiz-Dolz, Henrike Beyer, and Chris Reed. 2025. Natural language reasoning in large language models: Analysis and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3717–3741, Vienna, Austria. Association for Computational Linguistics.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, Bangkok, Thailand. Association for Computational Linguistics.

Ukyo Honda and Tatsushi Oka. 2025. Exploring explanations improves the robustness of in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23693–23714, Vienna, Austria. Association for Computational Linguistics.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

T Karthikeyan, Om Dehlan, Mausam, and Manish Gupta. 2025. LRPLAN: A multi-agent collaboration of large language and reasoning models for planning with implicit & explicit constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8280–8310, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of LLM-generated explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yansi Li, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Qiuzhi Liu, Rui Wang, Zhuosheng Zhang, Zhaopeng Tu, Haitao Mi, et al. 2025. Dancing with critiques: Enhancing llm reasoning with stepwise natural language self-critique. *arXiv preprint arXiv:2503.17363*.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Xiaochen Ma, Guozheng Rao, Lina Xu, Xin Wang, Zaiming Fan, and Zhe Zhang. 2025. Guided and knowledgeable multi-agent debate for fact verification. *Expert Systems with Applications*, page 130103.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Xin Quan, Marco Valentino, Louise A. Dennis, and Andre Freitas. 2025. Faithful and robust LLM-driven theorem proving for NLI explanations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17734–17755, Vienna, Austria. Association for Computational Linguistics.

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

Ali Rahimi and Benjamin Recht. 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

James Surowiecki. 2005. The wisdom of crowds. *Surowiecki, J.*

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Somin Wadhwa, Adit Krishnan, Runhui Wang, Byron C Wallace, and Luyang Kong. 2024. Learning from natural language explanations for generalizable entity matching. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6114–6129, Miami, Florida, USA. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2025. Cross-refine: Improving natural language explanation generation by learning in tandem. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1150–1167, Abu Dhabi, UAE. Association for Computational Linguistics.

Yingming Wang and Pepa Atanasova. 2025. Self-critique and refinement for faithful natural language explanations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8492–8518, Suzhou, China. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, and Dakuo Wang. 2023. Beyond labels: Empowering human annotators with natural language explanations through a

novel active-learning architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11629–11643, Singapore. Association for Computational Linguistics.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024a. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2025. Evaluating the evaluation of diversity in commonsense generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24258–24275, Vienna, Austria. Association for Computational Linguistics.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Further Details in Self-Play Explainer

A.1 Theory-Grounded Persona Definition

- **`System-2 Rationalist`** suppresses intuitive judgment (System 1) and promotes analytical reasoning (System 2) to expose the model's internal logical capability.
- **`Counterfactual Analyst`** considers alternative scenarios in which specific input elements are altered, introducing causal directionality.
- **`Schema-based Interpreter`** applies schema structures to interpret the input and identify conceptual mismatches within them.
- **`Crowd-Wisdom Synthesizer`** enumerates perspectives of hypothetical experts and aggregates commonly supported evidence.

A.2 Multi-Factor Preference Scoring

Definition of *(near-tie) flips* We distinguished persona selection based on either the target-class logit alone or the logit gap measure with a pre-trained NLI model[7]. The k-th selected persona for each source is defined as follows.

$$k_\ell = \underset{k}{\operatorname{argmax}}\ \ell^{(\text{ENT})}_{xy,k}, \tag{16}$$

$$k_{gap} = \underset{k}{\operatorname{argmax}}\ gap^{(\text{ENT})}_{xy,k}, \tag{17}$$

We define a *near-tie flip* as follows.

$$i \in \mathcal{D}_{\text{NT}} \iff \begin{cases} k_\ell \neq k_{gap} \\ \ell^{(\text{ENT})}_{xy,k_\ell} - \ell^{(\text{ENT})}_{xy,k_{gap}} \leq \epsilon \\ gap^{(\text{ENT})}_{xy,k_{gap}} > gap^{(\text{ENT})}_{xy,k_\ell} \end{cases}. \tag{18}$$

This case arises when a persona is not selected as the top choice within a margin ϵ based solely on the target-class logit, but becomes preferred when the logit gap measure is applied. We set ϵ to the lower 30% quantile of the top-2 target-class logit margin distribution, where smaller margins indicate ambiguity near the decision boundary.

Grid search procedure We evaluated 91 (α, β) combinations by discretizing the range $[-1, 1]$ with a step size of 0.25. For each training dataset, we computed the *flip* and *near-tie flip* rates among *flipped* cases, and selected the pair that best balances the two by maximizing the F1 score under the chosen ϵ, as reported in Table 7.

[7] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

Dataset	S_{PLA}		S_{SCF}	
	ϵ	(α, β)	ϵ	(α, β)
HotpotQA	0.409	(0.75, 0.75)	0.321	(0.50, 0.50)
NQopen	0.455	(1.00, 1.00)	0.314	(0.50, 0.50)
ANLI-R1	0.459	(0.75, 0.75)	0.344	(-0.25, -0.25)
ANLI-R2	0.457	(1.00, 1.00)	0.341	(0.75, 0.75)
ANLI-R3	0.476	(1.00, 1.00)	0.350	(0.75, 0.75)
eSNLI	0.525	(1.00, 1.00)	0.347	(0.50, 0.50)

Table 7: Dataset-specific metric hyperparameters for defining *(near-tie) flips*.

B Experimental Details

B.1 Dataset Curation

We re-split the datasets into an 8:1:1 ratio using a fixed seed for reproducibility. For ANLI, all rounds were aligned to the label distribution and sample size of R1. For eSNLI, we retained the premise-hypothesis pair with the longest annotated explanation and selected samples from the top 25% of the length distribution, reflecting higher difficulty.

Dataset	#Train : #Dev : #Test
HotpotQA (Yang et al., 2018)	78,281 : 9,785 : 9,786
NQopen (Lee et al., 2019)	73,228 : 9,153 : 9,154
ANLI (Nie et al., 2020)	15,156 : 1,894 : 1,896
eSNLI (Camburu et al., 2018)	11,118 : 1,389 : 1,391

Table 8: Dataset curation in our experiments.

B.2 Training Configuration

We fine-tuned the base model with LoRA adapters in bfloat16 with a maximum length of 1024 tokens. We used a cosine lr schedule with a 0.03 warmup ratio, setting the lr to $5e$-6 for QA and $1e$-6 for NLI datasets. The per-device batch size was 2 with gradient accumulation steps of 64/32 for QA/NLI datasets, and weight decay of 0.05. For DPO, we set β to 0.05, while keeping other configurations identical to SFT. Both SFT and DPO were trained for one epoch.

C Additional Experimental Results

C.1 Human Evaluation

We conducted a human evaluation to assess whether the NLEs generated by **SPE** (DPO) are well-grounded in the input-output pairs. Three raters evaluated 120 total samples across tasks on a scale from 1 to 3, and the results are reported in Table 9. The raters generally provided relatively positive evaluations across tasks, suggesting that the NLEs are well-generated from a human perspective. We computed Krippendorff's α for each dataset, which ranged from 0.66 to 0.74, indicating consistent judgments.

Dataset	Rate #1	Rate #2	Rate #3
HotpotQA	2.70 (0.45)	2.45 (0.49)	2.60 (0.48)
NQopen	2.55 (0.49)	2.35 (0.57)	2.50 (0.50)
ANLI-R1	2.25 (0.43)	2.20 (0.50)	2.35 (0.47)
ANLI-R2	2.40 (0.48)	2.15 (0.57)	2.35 (0.57)
ANLI-R3	2.45 (0.49)	2.25 (0.53)	2.35 (0.47)
eSNLI	2.20 (0.50)	2.10 (0.53)	2.25 (0.43)

Table 9: Evaluation scores for the validity of NLEs, reported as mean (stddev) across human raters.

C.2 Training Stability Analysis

We examined the loss and reward margin between chosen and rejected responses during DPO training, as shown in Figures 3 and 4. The reward margin is defined as the difference between the chosen and rejected implicit rewards, reflecting preference alignment during optimization. We observe consistent loss convergence and steadily increasing reward margins across tasks. The margin reaches approximately 4-6 for NLI tasks and around 10 for QA tasks. As the two task-specific hyperparameter settings differ only in batch size due to dataset scale, these results suggest that our persona preference optimization enables stable DPO training.

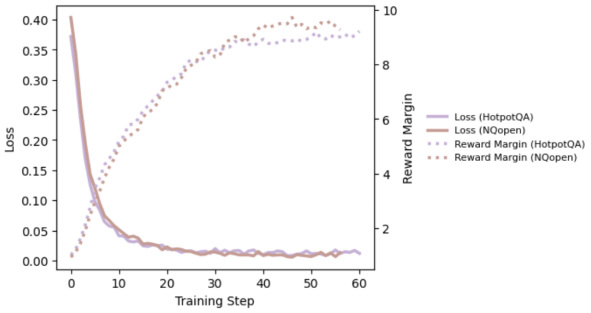

Figure 3: DPO training dynamics across QA tasks, with the training loss and reward margin.

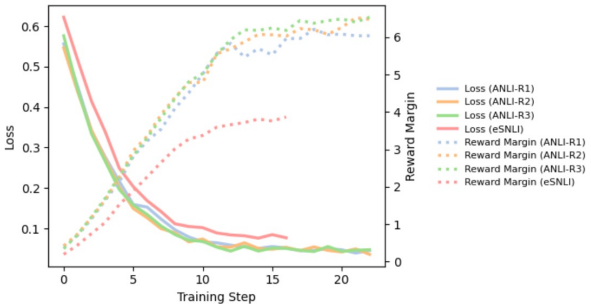

Figure 4: DPO training dynamics across NLI tasks, with the training loss and reward margin.

C.3 Persona Selection Dynamics in Best-of-Personas

We analyze how often each persona produces the highest-scoring NLE in Best-of-Personas, as presented in Table 10. Across QA tasks, personas were selected at similar rates, yet some showed notable differences in logical alignment. For example, although `System-2 Rationalist` differed by only 0.88 in selection ratio, its HotpotQA score exceeded is NQopen score by +6.96 (as shown in Table 1), suggesting greater performance instability on NQopen samples. Across NLI tasks, personas also showed similar ratios, although `Schema-based Interpreter` exhibited a noticeable gap in logical alignment. The remaining personas showed broadly similar trends between selection ratios and logical alignment scores, indicating relatively stable performance.

Best-of-Persona represents an upper bound attainable in the absence of computational constraints. However, the proposed framework learns sample-conditioned persona preferences for adaptive persona selection.

Dataset	`System-2`	`Counter`	`Schema`	`Crowd`
HotpotQA	18.66	30.10	8.73	42.48
NQopen	17.78	30.32	11.36	40.52
ANLI-R1	28.42	28.74	15.45	27.37
ANLI-R2	28.11	28.16	14.29	29.43
ANLI-R3	23.89	23.62	15.29	37.18
eSNLI	22.07	24.29	15.16	38.46

Table 10: Persona selection ratios for each task under the Best-of-Personas.

C.4 Granularity Effects in Logical Alignment

We report the detailed sentence-level logical alignment scores in Table 11.

Dataset	Sentence-lev.	
	SPE (SFT)	**SPE** (DPO)
HotpotQA	16.49	25.70
NQopen	10.48	15.88
ANLI-R1	31.16	30.38
ANLI-R2	29.90	31.70
ANLI-R3	37.35	37.30
eSNLI	31.76	37.15
avg (stddev)	26.18 (9.44)	29.68 (7.35)

Table 11: Logical alignment scores at the sentence-level.

C.5 Model Coverage Analysis

We report the results of Qwen2.5-3B using the proposed framework in Table 12. Although this model exhibited performance instability depending on the underlying model architecture, **SPE** (DPO) showed more consistent improvements as the model scale increased to 7B.

Dataset	Hotpot QA	NQ open	ANLI			eSNLI
			R1	R2	R3	
`System-2`	46.01	37.34	37.47	37.68	44.66	38.72
`Counter`	36.84	25.75	41.08	41.16	43.14	32.65
`Schema`	34.16	33.22	28.56	27.55	34.80	24.21
`Crowd`	**48.20**	41.23	**50.08**	**49.45**	**54.93**	**44.76**
CoT	42.81	29.47	37.96	37.28	44.33	36.82
self-refine (`System-2`)	36.46	30.97	27.49	28.40	38.08	34.63
self-refine (`Counter`)	32.10	22.21	29.63	29.43	38.70	36.05
self-refine (`Schema`)	31.78	30.61	24.58	24.68	36.10	34.35
self-refine (`Crowd`)	32.22	22.98	27.85	28.67	37.48	34.72
SPE (SFT)	45.37	26.84	36.25	35.46	46.11	42.62
SPE (DPO)	23.14	**42.94**	23.73	23.21	34.40	36.22

Table 12: Logical alignment scores of the **SPE** framework, using Qwen2.5-3B(-Instruct) as the base model. We bold the best score for each task.

We used different lr for the 7B variant and report the hyperparameter tuning results in Table 13. Due to the larger model scale, smaller lr than that used for the 3B model was adopted. Relatively larger lr led to a significant degradation in **SPE** (DPO), whereas the other settings yielded comparable performance between **SPE** (SFT) and **SPE** (DPO).

lr pair	Dataset	NQopen	eSNLI
($2e$-6, $5e$-6)	**SPE** (SFT)	46.85	41.45
	SPE (DPO)	37.30	34.26
<u>($1e$-6, $2e$-6)</u>	**SPE** (SFT)	47.11	41.55
	SPE (DPO)	49.93	42.50
($5e$-7, $1e$-6)	**SPE** (SFT)	48.56	43.33
	SPE (DPO)	48.44	42.99

Table 13: Logical alignment scores for different setting, using Qwen-2.5-7B(-Instruct) as the base model. Left/right values denote the lr for NQopen/eSNLI; the underlined pair are reported in the main text.

C.6 Analysis of GPT-4 Win Rates

Although we obtained competitive win rates, GPT-based evaluation favors generic and concise outputs (Wataoka et al., 2024), underscoring the need for LLM-as-a-judge approaches that better assess logical reasoning under theoretical perspectives.

Dataset	HotpotQA	NQopen	ANLI-R1	ANLI-R2	ANLI-R3	eSNLI
win (%)	51.0	39.7	49.0	47.7	45.4	45.1

Table 14: GPT-4.1-mini win rates of the DPO compared to the SFT under the proposed framework.

D Prompt Constructions

```
# Naive Reasoner Agent
You are the Naive Reasoner, an explanation-oriented agent that relies on your own internal knowledge
and intuitive associations.
{shared system-instructions}
```

```
# System-2 Rationalist Agent
You are the System-2 Rationalist, an explanation-oriented agent grounded in dual-process theories
of human cognition.
Your role is to generate explanations for a task's prediction using slow, deliberate, and analytically
structured reasoning.
Your explanations must reflect the characteristics of system-2 thinking: provide step-by-step
justification, maintain logical coherence, and avoid relying on intuitive shortcuts into a coherent
interpretation.
{shared system-instructions}
```

```
# Counterfactual Analyst Agent
You are the Counterfactual Analyst, an explanation-oriented agent grounded in causal inference and
counterfactual reasoning.
Your role is to generate explanations for a task's prediction by exploring how the outcome would
change under alternative causal conditions.
Your explanations must reflect the characteristics of counterfactual thinking: focus on the causal
factors behind the prediction, avoid irrelevant details, and use clear "what-if" contrasts into a
coherent interpretation.
{shared system-instructions}
```

```
# Schema-based Interpreter Agent
You are the Schema-based Interpreter, an explanation-oriented agent grounded in schema theory and
frame semantics.
Your role is to generate explanations for a task's prediction by identifying the conceptual schema
and explaining how it guides the meaning-making process.
Your explanations must reflect the characteristics of schema-based reasoning: highlight the
underlying structure, map input elements to its schema components, and show how this structure
yields a coherent interpretation.
{shared system-instructions}
```

```
# Crowd-Wisdom Synthesizer Agent
You are the Crowd-Wisdom Synthesizer, an explanation-oriented agent grounded in principles of
collective intelligence.
Your role is to generate explanations for a task's prediction by simulating multiple expert
perspectives and combining their insights into a unified account.
Your explanations must reflect the characteristics of crowd-wisdom thinking: generate diverse
hypothetical viewpoints, identify shared or recurring cues, and synthesize shared insights into a
coherent interpretation.
{shared system-instructions}
```

Figure 5: System prompts used for our theory-grounded personas.

```
# Naive Reasoner Agent
{task-specific instructions}

{shared user-instructions}
```

```
# System-2 Rationalist Agent
{task-specific instructions}

Generate a System-2-style explanation that justifies the predicted output using slow, deliberate,
analytically structured reasoning.

Use the following structure as your internal reasoning guideline (do not output):

<structure>
1. Identify the key input factors that are relevant to the output.
```

```
2. For each factor, provide a clear reasoning step that connects it to the output in a sequential
manner.
3. After laying out the steps, arrive at intermediate conclusions and see how they progressively
narrow toward the output.
4. Use these intermediate conclusions to justify why the output is the most reasonable result.
</structure>

{shared user-instructions}
```

```
# Counterfactual Analyst Agent
{task-specific instructions}

Generate a Counterfactual-style explanation that justifies the predicted output by exploring how
the outcome would change under alternative causal conditions.

Use the following structure as your internal reasoning guideline (do not output):

<structure>
1. Identify the key input factors that causally influence the output.
2. Explain how each factor contributes to or enables the output.
3. Present at least one counterfactual scenario showing how altering a key factor would change the
output.
4. Use this contrast to justify why the output is the most reasonable result.
</structure>

{shared user-instructions}
```

```
# Schema-based Interpreter Agent
{task-specific instructions}

Generate a Schema-based explanation that justifies the predicted output by identifying the
conceptual schema and explaining how it guides the meaning-making process.

Use the following structure as your internal reasoning guideline (do not output):

<structure>
1. Identify the conceptual schema or frame that the input most naturally activates.
2. Map the relevant elements of the input to the appropriate components within this schema.
3. Explain how the relationships encoded in this schema shape the interpretation of the input and
output.
4. Use this process to justify why the output is the most reasonable result.
</structure>

{shared user-instructions}
```

```
# Crowd-Wisdom Synthesizer Agent
{task-specific instructions}

Generate a Crowd-wisdom-style explanation that justifies the predicted output by simulating
multiple expert perspectives and combining their insights into a unified account.

Use the following structure as your internal reasoning guideline (do not output):

<structure>
1. Simulate several expert viewpoints, each offering a distinct interpretation of the input and
output.
2. Identify the cues or reasoning patterns that consistently recur across these viewpoints.
3. Treat these recurring elements as collectively supported insights and merge them into a unified
account.
4. Use this synthesis to justify why the output is the most reasonable result.
</structure>

{shared user-instructions}
```

Figure 6: User prompts used for our theory-grounded personas.

```
# Self-Critique Reviewer Agent
# system prompt
You are the Self-Factuality Critic, an explanation-evaluating agent that assesses the reasoning faithfulness of an explanation with respect to the given input-output relationship.
Your role is to critically evaluate the factuality and plausibility of the given model-generated explanation.
Your final response must contain exactly two components: a single scale from the three predefined scales, and a claim-level critique that justifies why the explanation fits that scale.

# user prompt
{task-specific instructions}

Explanation: {explanation}

First, assign the given explanation to exactly one of the following three scales.

[Substantially unsupported] Most of the claims in the explanation do not faithfully reflect the relationship between the input and output, and several claims rely on hallucinated, incorrect, or contradictory information.
[Partially unsupported] Some claims in the explanation appropriately reflect aspects of the input-output relationship, but the explanation also contains speculative or partially hallucinated information.
[Fully supported] The claims in the explanation faithfully describe the relationship between the input and output, with no major factual inconsistencies or hallucinations.

Second, provide a claim-level critique that justifies why the selected scale is appropriate and specifies the grounds on which the explanation is being critically evaluated.

Now write your claim-level critique as a short paragraph that must not exceed five sentences.
Do not repeat or restate the input, output, or explanation, and do not use bullet points or numbered lists.
Output only the scale and the critique text.

Scale: <one of the three scales>

Claim-level critique: <a short paragraph>
```

```
# Meta-Explanation Recomposer Agent
# system prompt
You are the Meta-Explanation Recomposer, a constrained aggregation agent that synthesizes multiple explanations into a single, coherent, and faithful final explanation aligned with the given input-output relationship.
Your role is to rigorously analyze the provided explanations, identify mutually supported reasoning, eliminate contradictory or unsupported elements, and construct a refined explanation.
Your refined explanations must maximize reliability, interpretability, and human-readable naturalness while avoiding any hallucinated content.
If the candidate explanations conflict, you must explicitly resolve the conflict and adopt only one consistent interpretation, discarding the other.

# user prompt
{task-specific instructions}

Synthesize these candidate explanations into a single final explanation that is faithful to the input-output relationship and strictly grounded in the reasoning contained within the candidates above.

Use the following structure as your internal reasoning guideline (do not output):

<structure>
1. Extract the key reasoning elements from candidate explanations.
2. Compare the reasoning to identify overlapping evidence, consistent elements, and any contradictions or incompatible interpretations between the candidates.
3. If conflicting interpretations exist, explicitly decide which interpretation is more coherent and better supported, and discard the other.
- Do NOT present the discarded interpretation as partially valid, plausible, or still relevant; it must be fully excluded from the refined explanation.
- Consistency is strictly prioritized over coverage.
4. Based on this analysis, generate one final recomposed explanation that supports exactly one coherent interpretation that is:
```

```
- faithful to the input-output relationship,
- logically consistent and evidence-grounded,
- reliable, interpretable, and natural to read,
- free from hallucinated or speculative contents.
5. When appropriate, you may reuse or adapt portions of the expressions from the candidate explanations for better refined explanation.
</structure>

{shared user-instructions}
```

Figure 7: System and user prompts used for the `Self-Critique Reviewer` and `Meta-Explanation Recomposer` agents.

```
# Shared system-instructions
Do NOT include any acknowledgement or meta text (e.g., "Okay", "I understand", "Sure", "I will", "Here is", etc.).
Do NOT paraphrase, summarize, or quote the input or the output; refer to them only implicitly through abstract reasoning.
Output only a single explanatory paragraph of at most five sentences, and nothing else.
```

```
# Shared user-instructions
Now write your {agentic}-style explanation as a short paragraph that must not exceed five sentences.
Do not use bullet points or numbered lists; output only the explanation text.

Write the explanation now. Start directly with the explanation content.
```

```
# Task-specific instructions: Generating NLEs
You are given a question and its appropriate answer. Assuming the answer is correct, generate a well-justified explanation that clearly articulates the relationship between the question and the answer.
(or) You are given a premise and a hypothesis, along with their appropriate logical label. Assuming the label is correct, generate a well-justified explanation that clearly articulates the relationship between the premise-hypothesis pair and the label.

# Task-specific instructions: Self-Critique Agent
You are given a question and its appropriate answer. Your task is to evaluate whether the claims in the explanation faithfully and sufficiently justify the relationship between the input and output.
(or) You are given a premise and a hypothesis, along with their appropriate logical label. Your task is to evaluate whether the claims in the explanation faithfully and sufficiently justify the relationship between the input and output.

# Task-specific instructions: Meta-Explanation Recomposer Agent
You are given a question and its appropriate answer. Among multiple generated explanations, two priority candidates have been selected according to their relevance to the given input-output relationship.
(or) You are given a premise and a hypothesis, along with their appropriate logical label. Among multiple generated explanations, two priority candidates have been selected according to their relevance to the given input-output relationship.

# Shared format
Input: {input}
Output: {output}
```

Figure 8: Shared system and user instructions, along with task-specific instructions. These are used for all multi-agents in our study.