Parser-Validated Minimal-Pair Preferences for DPO: A Case Study in Legal "Empathy"

Anonymous ACL submission

Abstract

Aligning LLMs to produce responses perceived as empathetic typically relies on costly human preference data and offers limited insight into which linguistic cues drive those preferences. We study legal question answering and introduce an automatic preference-data pipeline based on parser-validated *syntactic minimal pairs*. Grounded in linguistic accounts of perspective-taking, we generate rule-labeled minimal pairs along five dimensions (pronouns, voice, tense, polite imperatives, evaluative adverbs) and validate the intended contrast with dependency parsing (82.7% success). From 1,785 questions, we produce 7,378 minimal pairs and fine-tune three 7–8B model families (LLaMA-3, Mistral, Gemma) with DPO. Human evaluation (3,309 judgments, 35 raters) prefers DPO over an SFT baseline (68.8% vs. 31.2%, $p < 0.001$, h=0.40), robust to length controls. Feature validation shows voice is the dominant above-chance contributor (80%, h=0.64), while other edits are register-sensitive. Overall, parser-validated minimal pairs provide an interpretable, scalable route to preference optimization and identify which cues align with human judgments in-domain.

1 Introduction

The alignment of Large Language Models (LLMs) with human values represents one of the most critical challenges in AI safety and deployment (Ouyang et al., 2022; Bai et al., 2022a). As LLMs become the backbone of conversational AI systems, ensuring they can engage empathetically with users has emerged as a key alignment objective. However, empathy in AI systems can be understood through two distinct lenses: emotional empathy (feeling what others feel) and linguistic empathy—the grammatical encoding of perspective-taking and cognitive understanding (Kuno and Kaburaki, 1977). While most alignment research focuses on subjective emotional responses requiring human evaluation, we take a different approach by leveraging linguistically grounded perspective-taking cues that can be operationalized as parser-checkable feature controls.

Reinforcement Learning from Human Feedback (RLHF) has become the dominant approach for alignment (Ziegler et al., 2020; Kaufmann et al., 2023), but suffers from fundamental limitations: expensive human annotations, reward model training instabilities, and lack of interpretability. These challenges are particularly acute for subjective concepts like emotional empathy, where human annotators show low agreement and definitions vary across contexts (Casper et al., 2023).

Direct Preference Optimization (DPO) (Rafailov et al., 2023) simplifies alignment by eliminating the reward modeling step, instead learning directly from preference data. However, this approach inherits the fundamental bottleneck of RLHF: dependence on human-annotated preferences, which poses significant scalability and consistency challenges. We propose an alternative paradigm that reduces reliance on human preference labels for training by automatically generating preference pairs through systematic manipulation of linguistically motivated feature contrasts. Specifically, we leverage Kuno's theory of linguistic empathy (Kuno and Kaburaki, 1977; Kuno, 1987), which argues that grammatical choices systematically encode speaker perspective and alignment with discourse participants. This yields an explicit inventory of contrasts that can be instantiated and verified at the level of surface realization.

Scope of "meaning preservation." Throughout the paper, when we say that a transformation "preserves meaning," we mean that it aims to preserve *propositional (truth-conditional) content*. We do not aim to hold constant other aspects (e.g., register or discourse stance), which are instead part of the controlled contrast under study.

Our key insight is that certain grammatical patterns, such as voice alternation, pronoun usage, and tense selection, can be implemented as controlled *syntactic minimal pairs* in generated responses. Accordingly, we construct *minimal pairs* that differ in one explicitly targeted feature and validate the realization of that feature with dependency parsing. We then assign preference labels by the pre-specified feature target (i.e., which side instantiates the contrast), rather than by an LLM-as-judge pipeline.

As illustrated in Figure 1, we develop an automatic preference generation pipeline that:

1. creates minimal pairs through rule-based generators (e.g., active vs. passive voice),
2. validates them using dependency parsers to check that the intended feature contrast is realized, and
3. trains via DPO using these rule-defined labels, followed by human evaluation to test alignment with perceived empathy in-domain.

Our approach yields efficient learning of feature discrimination on the constructed minimal pairs across model families. Importantly, high discrimination accuracy on constructed pairs should be interpreted as learnability of the targeted contrast, not as a direct measure of empathetic generation quality. In our legal QA case study, we additionally conduct human evaluations to quantify which contrasts align with perceived empathy and to provide an empirical ranking of which linguistically motivated cues transfer in-domain.

2 Related Work

2.1 LLM alignment methods

Current LLM alignment methods face a fundamental tension between effectiveness and scalability. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2020) has become the dominant paradigm, using human preferences to train reward models that guide models toward helpful, honest, and harmless behavior (Bai et al., 2022a). However, RLHF suffers from well-documented limitations: reward hacking, distribution shift, prohibitive annotation costs, and annotator biases (Casper et al., 2023). These challenges are particularly acute for subjective qualities like emotional empathy, where definitions vary across contexts and inter-annotator agreement remains low (Kaufmann et al., 2023). Direct Preference Optimization (DPO) (Rafailov et al., 2023) addresses some technical limitations by eliminating the reward model and directly optimizing policies from preference data. While this simplifies training and improves stability, DPO inherits RLHF's core bottleneck: dependence on human-annotated preferences.

2.2 Automated preference generation

Recent work has explored automated preference generation to overcome the scalability barrier of human annotation. Constitutional AI (Bai et al., 2022b) generates feedback using predefined principles, allowing models to self-critique based on constitutional rules. Self-Instruct (Wang et al., 2023) enables models to generate their own training data through iterative instruction following and refinement. UltraFeedback (Cui et al., 2023) leverages GPT-4 to create multi-aspect preference judgments across various dimensions of response quality.

However, these approaches share critical limitations: (1) they rely on LLM-as-judge paradigms where AI systems make subjective quality assessments, potentially amplifying biases; (2) they lack theoretical grounding explaining *why* certain responses should be preferred; and (3) they typically do not provide a direct, feature-level linkage between the training label and a verifiable linguistic contrast, and often require separate validation against human judgments. In contrast, our approach grounds preferences in formal linguistic theory, uses parsers for objective verification, and validates alignment with human judgment through evaluation (80% acceptance for voice transformations).

2.3 Empathy modeling in NLP

Empathy modeling in NLP has gained attention through shared tasks and benchmarks (Hasan et al., 2023, 2024b,a). However, these approaches typically focus on emotional empathy, relying on subjective human annotations of empathetic responses and facing inter-annotator agreement challenges. Existing work largely treats empathy as an emergent property to be learned from labels rather than a phenomenon with systematic linguistic correlates (Kann, 2017). This conflation of emotional and linguistic empathy limits both the consistency of training signals and the interpretability of learned behaviors. Our work specifically targets linguistic empathy, objectively measurable through grammatical analysis.

2.4 Linguistic perspectives on empathy

While NLP approaches typically treat empathy as a holistic emotional quality requiring human judgment, linguistic research reveals that empathy is systematically expressed through grammatical choices. Kuno and Kaburaki's seminal work shows that linguistic empathy can be displayed via grammatical structure choice without altering propositional meaning (Kuno and Kaburaki, 1977). This linguistic empathy refers to the speaker's identification with participants in the described event, as reflected through syntactic choices (Kuno, 1987), namely, a formal property distinct from emotional empathy.

For instance, voice alternation shifts perspective without changing semantic content: structuring a sentence in active voice rather than passive voice (e.g., "John hit Mary" vs. "Mary was hit by John") changes the recipient of empathy by shifting *the camera angle* (Kuno and Kaburaki, 1977). Active voice centers the agent's perspective (John), while passive voice centers the patient's perspective (Mary). These are objectively verifiable grammatical properties, not subjective emotional assessments. Similarly, pronoun usage reveals psychological distance in discourse (Kuno, 1987). Choosing "he" or "she" instead of a proper noun indicates closer alignment with the referent (Kuroshima and Iwata, 2016). Recently, Kann (Kann, 2017) provided empirical validation for these linguistic theories by developing computational methods to measure perspective-marking features, finding significant correlations with psychological empathy scales.

Building on these theoretical foundations, we also incorporate evaluative adverbs based on Ernst's (Ernst, 2002) classification of agent-oriented adverbs. These adverbs (e.g., 'carefully,' 'thoughtfully,' 'thoroughly') express the speaker's assessment of how an agent performs an action, indicating cognitive effort and procedural care. While not direct empathy markers, we hypothesize that in professional contexts like legal discourse, explicit markers of deliberate consideration correlate with cognitive perspective-taking.

3 Method

3.1 Task formulation and approach

We formulate empathetic alignment as a preference learning task where responses exhibiting specific linguistic features associated with cognitive perspective-taking are preferred over those lacking them. Crucially, we target linguistic empathy (formal grammatical properties that encode perspective) rather than emotional empathy, enabling objective verification enabling parser-based verification of the targeted feature contrasts without requiring human-labeled preference data for training.

Given a legal question q, we generate preference pairs $(r_{accepted}, r_{rejected})$ that differ only in a single targeted feature while aiming to preserve propositional (truth-conditional) content. For each feature $f \in \{pronoun, voice, tense, politeness, evaluative\}$ and question q:

- $r^f_{accepted}$: response exhibiting feature f (objectively verifiable)
- $r^f_{rejected}$: response lacking feature f (objectively verifiable)

We then train models using Direct Preference Optimization (DPO) (Rafailov et al., 2023) to learn preferences for these perspective-encoding features. Given our generated dataset $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$, we optimize:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \quad (1)$$

following the standard DPO formulation. The high quality of our automatically generated contrast pairs—labeled by explicit feature targets and validated for their realization—provide clear supervision signals without requiring subjective human judgment or reward models.

3.2 Computational rules for linguistic empathy

We implement five computational rules that systematically transform linguistic features associated with perspective-taking and speaker stance. Each rule operationalizes a formal linguistic property that can be objectively verified through grammatical analysis:

Rule 1: Pronoun Replace full noun phrases with pronouns to reduce referential distance (Kuno, 1987). The rule identifies main subject positions and substitutes appropriate pronouns while preserving grammatical correctness.

✓ *"It must consider all evidence."*

× *"The tribunal must consider all evidence."*

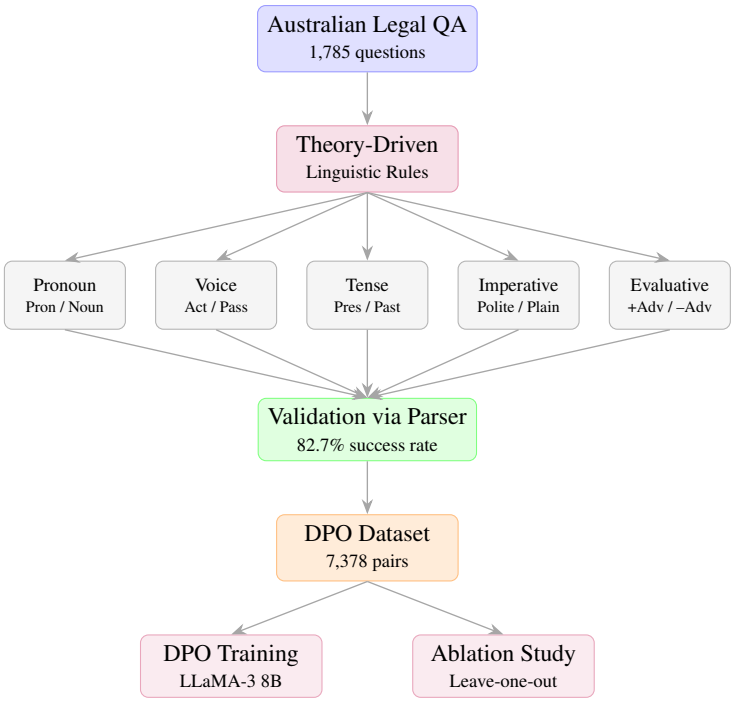


Figure 1: Automatic preference generation pipeline. Legal questions (1,785) are processed through five rule-based generators to create minimal pairs, validated by dependency parsers (82.7% success rate), producing 7,378 training pairs for DPO training.

Rule 2: Voice Convert between active and passive voice to shift perspective focus (Kuno and Kaburaki, 1977). The rule identifies transitive verbs and applies syntactic transformations—a formal grammatical operation verifiable through dependency parsing.

✓ *"The court reviews each application carefully."*
× *"Each application is reviewed carefully by the court."*

Rule 3: Tense Transform between present and past tense to manipulate temporal distance (Kann, 2017). The rule ensures consistent tense mapping throughout the response.

✓ *"The department processes requests promptly."*
× *"The department processed requests promptly."*

Rule 4: Politeness imperatives Add or remove politeness markers in imperatives (Brown and Levinson, 1987). The rule detects imperative constructions and systematically modifies politeness markers, a binary feature verifiable through syntactic analysis.

✓ *"Please submit your documents by Friday."*
× *"Submit your documents by Friday."*

Rule 5: Evaluative adverbs Insert evaluative adverbs from a predefined set indicating cognitive effort (Ernst, 2002). Presence/absence of specific lexical items from our defined set—objectively verifiable.

✓ *"The tribunal carefully examines each case."*
× *"The tribunal examines each case."*

3.3 Automatic preference generation pipeline

As illustrated in Figure 1, our pipeline implements these computational rules through two core components that enable fully automatic preference data generation with objective verification:

Rule-based generators For each computational rule, we develop specialized generators that produce minimal pairs differing only in the target linguistic feature. We use Claude-3-Haiku (Anthropic, 2024) to generate base responses, then apply deterministic rule-based transformations to create *accepted* and *rejected* variants. Crucially, the preference label is determined by the transformation rule—not by LLM judgment—ensuring rule-defined labels (i.e., the preferred side is determined by the targeted feature), rather than by an LLM judge. Detailed prompts are provided in the Appendix.

Dependency validators We implement rule-specific parsers using spaCy (Honnibal et al., 2020) to validate correct feature implementation through formal grammatical analysis. Pronoun validators

prevent ambiguity by preserving relative clause boundaries, while voice parsers detect auxiliary verbs and past participles to ensure grammatically correct active-passive conversions. Tense validation maintains consistency across finite verbs but preserves modals and infinitives where tense marking doesn't apply. For imperatives, our parser distinguishes genuine commands from declarative sentences, and for evaluative adverbs, we follow Ernst's classification (Ernst, 2009) to verify proper verb modification. This comprehensive validation achieves 82.7% success rate, with failed pairs regenerated up to three times using targeted feedback. Throughout, formatting and punctuation remain constant to prevent spurious learning signals. The validation process ensures that each pair instantiates the intended single-feature contrast under parser-based checks, with labels defined by the targeted feature.

4 Experiments

4.1 Dataset generation and processing

We apply our pipeline to 1,785 unique questions from the Open Australian Legal QA dataset (Butler, 2023). Due to varying complexity of linguistic transformations, we observe differential success rates across rules. From initial generation attempts, we successfully create 7,378 validated preference pairs: imperative (1,641 pairs), evaluative adverbs (1,557 pairs), tense (1,459 pairs), voice (1,445 pairs), and pronoun substitution (1,276 pairs). The higher success rates for imperative and evaluative rules reflect their straightforward transformations, while pronoun substitution proves most challenging due to complex co-reference resolution requirements.

To validate cross-architecture robustness, we train three model architectures: Meta-Llama-3-8B-Instruct (Dubey et al., 2024; Meta, 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023; Mistral AI, 2023), and Gemma-7B-IT (Gemma Team, 2024; Google, 2024). For ablation studies requiring controlled comparisons, we use a subset of 519 questions where all five rules successfully generated valid pairs, yielding 2,595 pairs (519 $\times$ 5 rules). Each leave-one-out condition contains 2,076 pairs (519 $\times$ 4 rules).

4.2 Training configuration

We fine-tune three model architectures (LLaMA-3 8B Instruct, Mistral-7B-Instruct, and Gemma-7B-IT) using the TRL library's DPO implementation. Our configuration employs LoRA (Hu et al., 2021) (r=16, α=32, dropout=0.05) for parameter-efficient training with batch size 16 (4 $\times$ 4 gradient accumulation), learning rate 5e-5 with cosine scheduling, and β=0.1. We implement early stopping with patience of 2 evaluations (every 10 steps) to prevent overfitting. All experiments run on a single NVIDIA A100 80GB GPU. Training dynamics are shown in Appendix F.

Model	Steps to 100%	Final Margin
Mistral-7B	50	15.26
LLaMA-3 8B	75	16.22
Gemma-7B-IT	100	13.95

Table 1: Cross-architecture convergence. All models achieve perfect preference discrimination within 100 steps.

5 Results

5.1 Training and Cross-Architecture Validation

We first verify that our linguistically-grounded minimal pairs provide learnable supervision signals across different model architectures. Training three models (LLaMA-3 8B, Mistral-7B, and Gemma-7B-IT) on 7,378 pairs, all achieved 100% preference discrimination within 50–100 steps with consistent margins (Table 1). This cross-architecture consistency confirms that our linguistic rules provide clear, model-agnostic signals rather than exploiting architecture-specific biases.

The 100% accuracy requires careful interpretation. This reflects *feature discrimination*—the ability to distinguish presence from absence of specific grammatical markers—not empathetic generation quality. The question of whether these learned preferences translate to human-perceived empathy requires direct evaluation.

5.2 Human Evaluation: DPO vs. SFT

To answer this question, we conducted an IRB-approved study comparing DPO-aligned models against a supervised fine-tuning (SFT) baseline trained on the same accepted responses. Qualified MTurk workers ($\geq$97% approval, $\geq$1000 HITs, US location) judged which of two responses to 50 legal questions was "more empathetic." Each item received 16 judgments, yielding 784 valid responses after quality control.

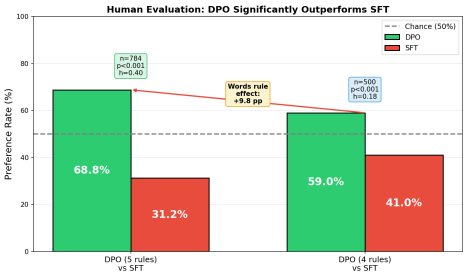


Figure 2: Human evaluation: DPO significantly outperforms SFT (68.8% vs. 31.2%, $p < 0.001$, Cohen's h=0.40).

The results were clear: DPO responses were preferred in 68.8% of judgments versus 31.2% for SFT ($p < 0.001$, Cohen's h=0.40; Figure 2). This medium-sized effect demonstrates that preference learning on linguistic features produces responses humans perceive as meaningfully more empathetic than supervised learning alone.

5.3 Controlling for Length Bias

One potential confound is response length: if DPO models simply generate longer responses, annotators might prefer them regardless of empathy. We tested this through correlation analysis and length-matched comparisons.

The correlation between length difference and DPO preference was weak and non-significant (r=0.178, p=0.222; Figure 3, left). More directly, in pairs where responses differed by only 0–5 words, DPO was still preferred in 58.2% of cases (Figure 3, right). Effect decomposition confirmed that quality contributes +10.5 percentage points while length contributes only −0.7pp. The DPO advantage reflects response quality, not verbosity.

5.4 Feature-Level Validation

Having established that DPO outperforms SFT overall, we next ask: which of our five linguistic rules actually drive perceived empathy? Using the Communicative Empathy framework (Buechel et al., 2018), nine raters evaluated 75 accepted/rejected training pairs across three dimensions (Understanding, Supportiveness, Helpfulness), yielding 2,025 judgments.

The results revealed striking asymmetry (Figure 4). Voice transformations achieved 80.0% acceptance (h=0.64, $p < 0.001$), consistent across all three empathy dimensions. The remaining rules fell at or below chance: Imperative (36.8%), Evaluative adverbs (19.8%), Tense (17.8%), Pronoun

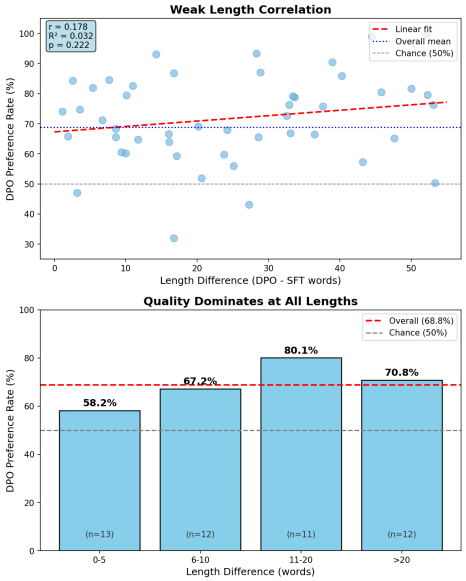


Figure 3: Length bias analysis. Top: Weak correlation (r=0.178, p=0.222). Bottom: DPO preferred even in length-matched pairs (58.2% at 0–5 word gap).

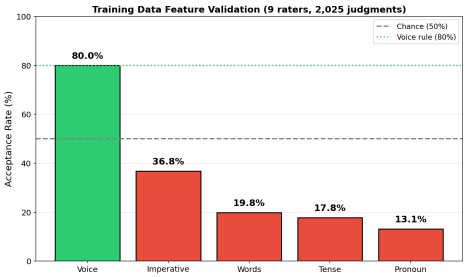


Figure 4: Feature validation (n=2,025). Only voice achieved above-chance acceptance (80%); other features conflict with legal register norms.

(13.1%). Voice emerges as the dominant feature for perceived empathy in legal domain.

5.5 Feature Ablation

The low individual acceptance rates raise a puzzle: if four of five features fail validation, why does the full model outperform SFT so substantially? We hypothesized that features might contribute synergistically even when individually weak.

To test this, we trained a model without the evaluative adverbs rule and repeated our human evaluation (n=500). Removing this rule reduced DPO preference from 68.8% to 59.0% (−9.8pp, p=0.033, Cohen's h=0.18; see Appendix E for details). Features that fail isolated validation can still contribute meaningfully in combination.

5.6 Training Dynamics Ablation

Finally, we examined how individual rules affect training dynamics through leave-one-out ablation on a balanced 519-question subset.

Removing voice produced a surprising result: margins *increased* 2.4× (21.32 → 51.43) while maintaining perfect accuracy. This contrasts sharply with voice's strong human validation (80%). Removing tense had the opposite effect: accuracy degraded to 99.5% with the lowest margins (18.57). Other rules (pronoun, politeness, evaluative) showed minimal impact on training dynamics, with margins ranging from 19.47 to 20.49.

This apparent paradox (i.e., voice is the strongest human-validated signal yet yields lower training margins when included) requires interpretation. This contrast suggests that optimization difficulty (as reflected by training margins) and human-perceived desirability need not align, especially when a feature is pragmatically appropriate only in a subset of legal contexts. In other words, margins primarily reflect how separable the constructed preference signal is under the model and dataset, whereas human judgments reflect context-sensitive pragmatic appropriateness. We revisit this discrepancy in the Discussion.

6 Discussion

6.1 From Black-Box Preferences to Interpretable Alignment

Traditional RLHF operates as a black box: annotators judge which response is "better," but the model cannot identify *which features* drive this preference. When thousands of subjective judgments coalesce into model behavior, the learned patterns resist interpretation.

Our approach inverts this paradigm. By decomposing "empathy" into five linguistically grounded computational rules, we create preferences with rule-defined labels tied to an explicit, parser-checkable contrast. When our model prefers response A over B, we can trace this to a specific grammatical transformation: A uses active voice while B uses passive, or A includes evaluative adverbs while B omits them.

This transparency enabled discoveries invisible to conventional methods: the dominance of voice (80% human acceptance), the domain-specificity of other features, and the synergistic effects that emerge only in combination.

6.2 Comparison with Automated Preference Methods

Recent work has explored automated preference generation to overcome annotation costs. Constitutional AI (Bai et al., 2022b), Self-Instruct (Wang et al., 2023), and UltraFeedback (Cui et al., 2023) use LLM-as-judge paradigms where models evaluate response quality. While scalable, these approaches share critical limitations our method addresses.

First, existing methods lack theoretical grounding explaining *why* certain responses should be preferred. Our preferences derive from formal linguistic theories of perspective-taking, where each grammatical choice encodes documented effects on speaker-referent alignment.

Second, LLM-generated preferences cannot be objectively verified. Our dependency parsers confirm the realization of the targeted contrast, providing verification independent of any LLM-as-judge preference labeling.

Third, prior work assumes LLM preferences approximate human perception without validation. We validate directly: voice achieves 80% human acceptance (h=0.64), confirming theoretical predictions, while also revealing that other features require domain adaptation.

6.3 Voice as the Dominant Signal

Human evaluation revealed striking asymmetry: voice achieved 80% acceptance while other features fell below chance in legal domain. This validates Kuno and Kaburaki's (1977) central claim that voice alternation encodes speaker perspective. Active voice positions the speaker closer to the agent, while passive voice distances the speaker. This "camera angle" shift is perceived as empathy.

Yet removing voice during training *increased* margins 2.4×. This apparent paradox reflects domain-specific pragmatics: in legal writing, passive voice serves objectivity ("The evidence was examined"), agent backgrounding ("Damages were awarded"), and precedent citation ("This principle was established in Smith v. Jones"). Our training data thus contained inconsistent signals. The 80% acceptance reflects the general preference for active voice; the 20% represents legal contexts where passive is appropriate.

This finding has implications for alignment: universal linguistic preferences require domain adaptation. Future systems could implement context-

aware weighting based on register and communicative goals.

6.4 Synergistic Feature Effects

Evaluative adverbs achieved only 19.8% individual acceptance, yet removing this rule reduced overall preference from 68.8% to 59.0%. How can a weak feature contribute to success?

We hypothesize that linguistic empathy is perceived holistically rather than through individual markers. A response combining active voice, present tense, and evaluative adverbs creates a cumulative impression of speaker engagement, even if each feature alone is insufficient. This has methodological implications: isolated feature validation may underestimate contributions that emerge only in combination. Our ablation design captures these interaction effects.

7 Limitations and Future Work

Several limitations warrant consideration. First, four of five rules conflict with legal register norms, and cross-domain evaluation is ongoing. Second, our single-sentence minimal pairs cannot capture discourse-level empathy patterns. Third, we operationalized only five computational rules grounded in linguistic empathy theory; additional phenomena (deixis, evidentiality, modal verbs) remain unexplored. Fourth, our English-only implementation requires cross-linguistic validation. Finally, we measured perceived empathy rather than downstream outcomes such as user satisfaction or therapeutic alliance.

8 Conclusion

We introduced a parser-validated pipeline for synthesizing preference data as syntactic minimal pairs and applying DPO with substantially reduced reliance on human-labeled preference data for training. In a case study on legal question answering, we operationalized five linguistically motivated feature contrasts, validated transformations with dependency parsing, and produced thousands of preference pairs for training.

Human evaluation shows that DPO trained on these minimal-pair preferences is preferred over an SFT baseline (68.8% vs. 31.2%). Feature-level validation reveals strong asymmetry: voice alternation is the dominant above-chance contributor to perceived empathy in this legal setting (80% acceptance), while other edits appear register-sensitive and may be dispreferred in isolation. Together, these findings suggest that minimal-pair preferences provide learnable, interpretable supervision for preference optimization, but the cues that transfer to perceived empathy depend on domain pragmatics rather than purely on grammatical form.

Our core contribution is interpretability: each preference pair is linked to a specific, verifiable grammatical contrast, enabling targeted analysis of what models learn and why certain cues succeed or fail in-domain. Future work will extend the pipeline to additional domains and languages and expand the feature inventory to discourse-level phenomena (e.g., deixis, evidentiality, modality) that may better capture context-sensitive perspective-taking.

Ethical Statement

This work studies linguistic cues related to perspective-taking and their association with perceived empathy in a legal question answering setting; we do not claim that grammatical manipulation yields genuine emotional understanding or legal advice. Human evaluation was conducted under IRB approval with informed consent and compensation at an estimated $16.50/hour. We use publicly available legal questions and will release code and derived artifacts for reproducibility, subject to dataset and model license constraints.

Acknowledgments

We thank Claude 4.5 for assistance with grammatical proofreading and for suggestions that improved the visual presentation of several figures.

References

Anthropic. 2024. Claude 3 haiku. https://www.anthropic.com/claude. Accessed: 2024-12-01.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*.

Umar Butler. 2023. Open australian legal qa.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Thomas Ernst. 2002. *The Syntax of Adjuncts*. Cambridge University Press, Cambridge.

Thomas Ernst. 2009. Speaker-oriented adverbs. *Natural Language & Linguistic Theory*, 27(3):497–544.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Google. 2024. google/gemma-7b-it: Model card. https://huggingface.co/google/gemma-7b-it. Accessed 2026-01-03.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024a. LLM-GEm: Large Language Model-Guided Prediction of People's Empathy Levels towards Newspaper Article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2215–2231, St. Julian's, Malta. Association for Computational Linguistics.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, Susannah Soon, and Shafin Rahman. 2023. Curtin OCAI at WASSA 2023 Empathy, Emotion and Personality Shared Task: Demographic-Aware Prediction Using Multiple Transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 536–541, Toronto, Canada. Association for Computational Linguistics.

Md Rakibul Hasan, Md Zakir Hossain, Aneesh Krishna, Shafin Rahman, and Tom Gedeon. 2024b. Thesis Proposal: Detecting Empathy Using Multimodal Language Model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 338–349, St. Julian's, Malta. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Theresa Kann. 2017. *Measuring linguistic empathy: An experimental approach to connecting linguistic and social psychological notions of empathy*. Ph.D. thesis, University of California, Los Angeles.

Timo Kaufmann, Sasha Ball, Jacob Beck, Eyke Hüllermeier, and Frauke Kreuter. 2023. On the challenges and practices of reinforcement learning from real human feedback. In *Proceedings of the ICML Workshop on Challenges in Deployable Generative AI*.

Susumu Kuno. 1987. *Functional syntax: Anaphora, discourse and empathy*. University of Chicago Press, Chicago.

Susumu Kuno and Etsuko Kaburaki. 1977. Empathy and syntax. *Linguistic Inquiry*, 8(4):627–672.

Satomi Kuroshima and Noriko Iwata. 2016. On displaying empathy: Dilemma, category, and experience. *Research on Language and Social Interaction*, 49(2):92–110.

Meta. 2024. meta-llama/meta-llama-3-8b-instruct: Model card. https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct. Accessed 2026-01-03.

Mistral AI. 2023. mistralai/mistral-7b-instruct-v0.2: Model card. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2. Accessed 2026-01-03.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Appendices A–D detail our automatic preference generation pipeline; Appendix E provides human evaluation methodology; and Appendix F provides training details.

A Complete Generator Prompts

We present the full implementation of our rule-specific generators, including all prompts, transformation logic, and fallback mechanisms.

A.1 Pronoun Substitution Generator

Stage 1: Base Response Generation

System: You answer legal questions with single sentences using full noun phrases.

User Prompt:
Answer this question with ONE sentence using FULL NOUN PHRASES.

CRITICAL: Do NOT use any pronouns (it, they, he, she, this, that).
Always use the complete noun phrase.

Examples:

✓ "The immigration tribunal must consider all evidence that the applicant submits."
✗ "It must consider..." (NO! Use "The tribunal must consider...")

Question: `{question}`

Stage 2: Pronoun Transformation

Transformation Rules:

- Replace ONLY the MAIN SUBJECT (before relative clause)
- KEEP relative clauses (who/that/which...) intact
- "The tribunal" → "It"
- "The person/applicant" → "They"
- "Individuals" → "They"

Example:
"The tribunal that reviews immigration cases must consider all evidence."
→ "It that reviews immigration cases must consider all evidence."

Stage 3: Fallback Patterns

Regex patterns (if LLM fails):

With relative clauses:

- `The (tribunal|court|board) + (who|that|which)...` → `It`
- `The (applicant|person) + (who|that|which)...` → `They`

Simple patterns:

- `The tribunal` → `It`
- `The court` → `It`
- `The applicant` → `They`

A.2 Voice Transformation Generator

Stage 1: Active Voice Generation

System: You answer legal questions with single sentences in active voice.

Requirements:

- Subject performs action (S + V + O structure)
- Include clear object for passive transformation
- Avoid "is/was/been + past participle"

Good examples:

✓ "The tribunal must review all submitted documents."
✗ "Documents are reviewed by the tribunal." (passive)

Stage 2: Passive Transformation

Transformation patterns:

- Basic: "The court decides cases" → "Cases are decided by the court"
- Modal: "The court must decide" → "It must be decided by the court"
- Continuous: "The court is reviewing" → "It is being reviewed by the court"

A.3 Tense Mapping Generator

Present Tense Generation

System: Describe what tribunals ACTUALLY DO in present tense.

Requirements:

- Use present ACTION verbs (reviews, processes, decides)
- AVOID modals (can, must, may) - describe actual actions

Tense mappings:
is→was, are→were, has→had, reviews→reviewed, decides→decided

A.4 Politeness Insertion Generator

Imperative Generation & Transformation

Stage 1: Generate plain imperative (no politeness markers)
Example: "Submit your appeal to the tribunal within 30 days."

Stage 2: Add politeness marker
- Formal contexts: Add "Kindly"
- General: Add "Please"
Example: "Submit your appeal" → "Please submit your appeal"

A.5 Evaluative Adverb Generator

Plain Statement & Adverb Insertion

Stage 1: Generate without evaluative adverbs
Example: "The tribunal considers all submitted evidence."

Stage 2: Insert appropriate adverb
Approved adverbs (agent-oriented, following Ernst's classification):

- Manner: carefully, thoughtfully, thoroughly, deliberately
- Procedural: properly, appropriately, correctly, duly

- Evaluative: reasonably, sensibly, fairly, wisely
- Professional: professionally, competently, efficiently, effectively

Example: "The tribunal considers" → "The tribunal carefully considers"

B Validation Algorithms

B.1 Base Validation Framework

Algorithm 1 Base Rule Validation

Require: Accepted R_a, Rejected R_r, Rule T
Ensure: Valid (bool), Diagnostics
$docA \leftarrow$ SpacyParse(R_a)
$docR \leftarrow$ SpacyParse(R_r)
if SentCount($docA$) $\neq 1$ OR SentCount($docR$) $\neq 1$ **then**
  return False, "Not single sentence"
end if
if WordCount($docA$) < 10 OR WordCount($docA$) > 40 **then**
  return False, "Length violation"
end if
$result \leftarrow$ ValidateRule($docA$, $docR$, T)
return $result$

B.2 Pronoun Validation

Algorithm 2 Pronoun Rule Validation

Require: Parsed $docA$, $docR$
Ensure: Validation result
$hasRelClause \leftarrow$ CheckRelativeClauses($docA$, $docR$)
$subjA \leftarrow$ ExtractMainSubjects($docA$)
$subjR \leftarrow$ ExtractMainSubjects($docR$)
$pronouns \leftarrow$ CountPronouns($subjA$)
$nouns \leftarrow$ CountNouns($subjR$)
if $pronouns = 0$ **then**
  return False, "No pronouns in accepted"
end if
if $nouns = 0$ **then**
  return False, "No nouns in rejected"
end if
if $hasRelClause$ AND NOT PreservedRelClause($docA$, $docR$) **then**
  return False, "Relative clause altered"
end if
return True, {pronouns: $pronouns$, nouns: $nouns$}

B.3 Other Validation Algorithms

[Similar compact format for Voice, Tense, Politeness, and Evaluative validations]

C Results and Analysis

C.1 Generation Success Rates

C.2 Failure Analysis

Common Failure Patterns

Pronoun (29.1% failure rate):
- Complex relative clauses with multiple referents (35%)
- Nested sentence structures (28%)

Table 2: Detailed generation statistics (7,378/8,925 total pairs = 82.7%)

Rule	Pairs	Success	1st try	Avg.
Politeness	1,641	91.2%	87.2%	1.2
Evaluative	1,557	86.5%	79.3%	1.4
Tense	1,459	81.1%	69.1%	1.7
Voice	1,445	80.3%	68.4%	1.8
Pronoun	1,276	70.9%	55.7%	2.1

- Ambiguous antecedents (22%)
- Coordination issues (15%)

Voice (19.7% failure rate):
- "Be + noun" copular structures from listing questions (40%)
- "Who + be" questions producing identity statements (17%)
- "Who represented" questions with awkward passivization (11%)
- Already passive constructions (9%)
- Other intransitive/stative verbs (23%)

Tense (18.9% failure rate):
- Mixed tense requirements in legal contexts (38%)
- Perfect aspect preservation issues (29%)
- Modal verb interference (24%)
- Conditional statements (9%)

Politeness (8.8% failure rate):
- Non-imperative structures (52%)
- Embedded commands (31%)
- Indirect speech acts (17%)

Evaluative (13.5% failure rate):
- Adverb placement conflicts (44%)
- Multiple verb phrases (35%)
- Auxiliary verb complications (21%)

C.3 Detailed Voice Failure Analysis

We analyzed 35 voice transformation validation failures in detail. The primary cause is that certain question types inherently elicit responses with copular or intransitive verbs, which cannot undergo active-passive transformation.

Table 3: Voice validation failure breakdown by question pattern (n=35)

Question Pattern	Count	%
"What was/were the X" (listing/be+noun)	14	40%
"Who was/were" (identity statement)	6	17%
"Who represented" (awkward passive)	4	11%
"Can X be" (already passive)	3	9%
Other (intransitive/stative)	8	23%

D Examples

D.1 Successful Transformations

Pronoun - Natural Relative Clause

Q: "What about individuals who have already filed an appeal?"

Base: "Individuals who have already filed an appeal must wait for the tribunal's decision before taking further action."

Transformed: "They who have already filed an appeal must wait for the tribunal's decision before taking further action."

✓ Relative clause preserved, only main subject replaced

Voice - Modal Handling

Q: "Must the court consider new evidence?"

Active: "The court must carefully examine all new evidence."

Passive: "All new evidence must be carefully examined by the court."

✓ Correct passive with modal preserved

D.2 Failure Cases

Pronoun - Coordination

Q: "Can the person who filed the complaint and the witness who supports it both appear?"

Base: "The person who filed the complaint and the witness who supports the complaint may both appear."

Failed: "They who filed the complaint and they who supports it may both appear."

✗ Double pronoun creates ambiguity; coordinated subjects require special handling

Voice - "What were the X" (Be + Noun)

Q: "What were the two issues in the appeal case of Millar v Commissioner of Taxation [2016] FCAFC 94?"

Expected answer structure: "The two issues were X and Y." (copular verb)

Problem: The verb "were" links subject to noun complement—no transitive action verb exists to passivize.

✗ Copular verbs cannot undergo active-passive transformation

Voice - "Who were the X" (Identity Statement)

Q: "Who were the parties involved in the case Kirby v Centro Properties Limited (No 2) [2011] FCA 1144?"

Expected answer: "The parties were X and Y." (be + noun predicate)

Problem: Identity statements use "be" as a linking verb, not a transitive action verb.

✗ No action verb to convert to passive voice

Voice - "Who represented" (Awkward Passive)

Q: "Who represented the applicant and the respondent in the case of CCA Beverages v Commissioner of Taxation [1995] FCA 980?"

Active: "Mr. Smith represented the applicant and Ms. Jones represented the respondent."

Attempted passive: "The applicant was represented by Mr. Smith and the respondent was represented by Ms. Jones."

Problem: While grammatically possible, passive construction obscures the answer to "who" questions.

✗ Semantic mismatch: passive form doesn't naturally answer "who" questions

Voice - Intransitive

Q: "When does the tribunal meet?"

Active: "The tribunal meets every Tuesday."

Failed attempts:

- "Every Tuesday is met by the tribunal" (nonsensical)
- "Meetings are held by the tribunal" (verb changed)

✗ No valid passive for intransitive "meet"

Voice - Already Passive Construction

Q: "In the context of New South Wales law, can a company be bound by a contract independently of the Corporations Act?"

Problem: Question already contains passive voice ("be bound"). Generating an active-voice answer and converting it back creates circular transformations.

✗ Question structure primes passive response

Tense - "When did" Questions

Q: "When did the Health Insurance Act amendment commence?"

Expected answer: "The amendment commenced on January 1, 2020." (past tense)

Problem: The question already uses past tense ("did"), naturally eliciting a past-tense response. Present-to-past transformation becomes meaningless.

✗ Question tense determines answer tense

Evaluative - Definition Questions

Q: "What is the definition of an RSE as per the Financial Sector (Collection of Data) reporting standard?"

Expected answer: "An RSE is defined as..." (stative/definitional)

Problem: Definition responses lack action verbs that evaluative adverbs can modify. Adding "carefully" or "thoroughly" to "is defined as" is semantically odd.

✗ Stative verbs incompatible with manner adverbs

E Human Evaluation Details

This appendix provides comprehensive details of our human evaluation studies, including platform specifications, participant qualifications, quality control measures, and complete study designs.

E.1 Platform and Recruitment

Platform All studies were conducted using Google Forms distributed via Amazon Mechanical Turk (MTurk). Figure 5 shows our HIT configuration.

Compensation Participants received $5.50 per HIT (estimated 20 minutes completion time), corresponding to an hourly rate of $16.50.

Participant Qualifications

- Location: United States only
- Language: Native English speakers
- MTurk approval rate: $\geq$97%
- Completed HITs: $\geq$1,000

Consent All participants confirmed: "I am 18+ and consent to participate" before proceeding.

E.2 Quality Control

Attention Checks Each study included mandatory attention check questions to ensure data quality:

> *"Attention check: To show you are reading, please select option B."*

Participants who failed attention checks were excluded from analysis and their HITs were rejected.

Response Order Randomization For all A/B comparisons, the presentation order of responses was randomized to prevent position bias. The `true_label` field in our data files tracks which response (A or B) corresponds to each condition.

E.3 Study 1: DPO vs. SFT Comparison

Objective Evaluate whether DPO-aligned responses are preferred over supervised fine-tuning (SFT) baseline responses.

Materials

- Response pairs: 50 held-out legal questions
- Conditions: DPO (5 rules) vs. SFT baseline

Procedure Participants viewed pairs of responses (labeled A and B) to legal questions and selected which response was "MORE EMPATHETIC overall."

Sample Size

- Questions: 50
- Judgments per question: 16
- Total judgments: 800 (784 after quality control)

Results DPO responses preferred 68.8% vs. SFT 31.2% (exact binomial $p < 0.001$, Cohen's h=0.40).

E.4 Study 2: Construct Validity

Objective Validate that our theoretical constructs (minimal pair transformations) translate to perceived empathy differences across multiple dimensions.

Materials

- Response pairs: 75 accepted/rejected minimal pairs
- Rules covered: pronoun, voice, tense, politeness, evaluative adverbs

Dimensions Based on the Communicative Empathy framework (Buechel et al., 2018):

1. **Understanding** (Perspective-taking)
2. **Supportiveness** (Affective concern)
3. **Helpfulness** (Action-oriented guidance)

Procedure Each of 75 pairs was evaluated on all 3 dimensions. Questions were labeled with dimension (e.g., "[Understanding] What was the outcome...").

Sample Size

- Pairs: 75, Dimensions: 3, Questions per rater: 225
- Raters: 9, Total judgments: 2,025

Results Voice transformations achieved 80% acceptance (h=0.64, p <0.001); other rules showed lower acceptance (13–37%).

E.5 Study 3: Length Bias Analysis

Objective Isolate the contribution of the evaluative adverbs rule from potential length confounds.

Materials

- Response pairs: 50 (same questions as Study 1)
- Conditions: DPO (4 rules, evaluative adverbs removed) vs. SFT

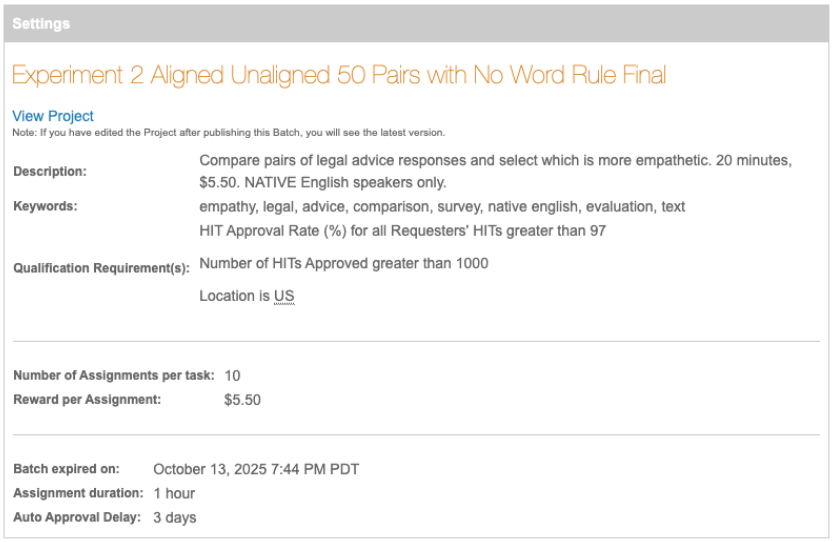


Figure 5: MTurk HIT settings: qualification requirements (US, ≥97% approval, ≥1000 HITs), compensation ($5.50), and assignment duration (1 hour).

Procedure Identical to Study 1, but using a DPO model trained without the evaluative adverbs rule.

Sample Size 50 questions, 500 total judgments.

Results Removing the evaluative adverbs rule reduced DPO preference from 68.8% to 59.0% (−9.8pp, p=0.033, Cohen's h=0.18). Effect decomposition reveals: Quality (direct) +10.5pp, Length (mediated) −0.7pp (Figure 6).

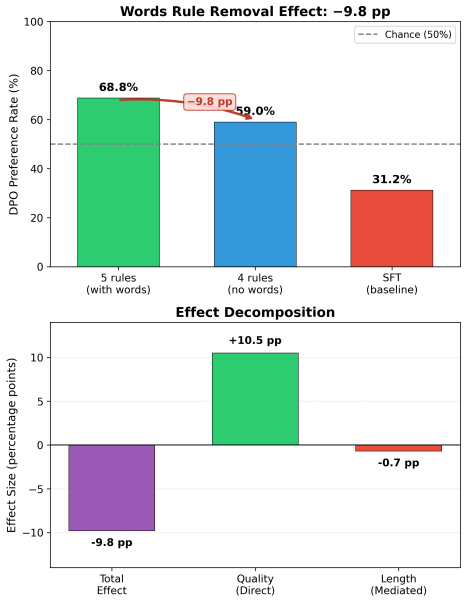


Figure 6: Words rule removal effect (−9.8pp) and decomposition: quality (+10.5pp) vs. length (−0.7pp).

E.6 Summary

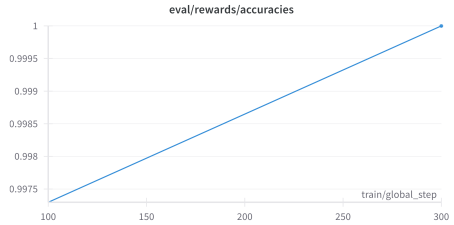
Table 4 summarizes all human evaluation studies.

Study	Pairs	Judg.	Raters
1: DPO vs SFT	50	784	16/q
2: Validity	225	2,025	9
3: Length	50	500	–
Total	–	**3,309**	**35**

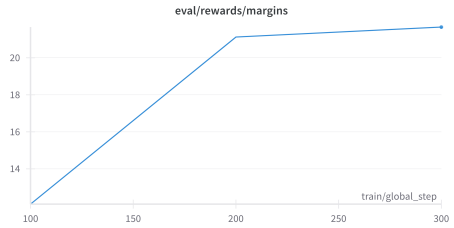
Table 4: Summary of human evaluation studies.

F Training Details

Figure 7 shows evaluation metrics for LLaMA-3 8B; similar patterns were observed for other models.



(a) Accuracy



(b) Margin

(c) Loss

Figure 7: Evaluation metrics during DPO training (LLaMA-3 8B).