Human-Like Decision Modeling With Small Language Models

Eunjin Yun
University of Texas at Austin
ej.yun@utexas.edu

YongGeon Lee
University of Texas at Austin
yg910524@utexas.edu

Abstract

Recently, large language models (LLMs) have served as computational models for human decision-making. However, existing research often depends on very large models and extensive behavioral datasets, which limits accessibility and interpretability. Here, we investigate whether a smaller LLM (2–7B parameters) can replicate human odd-one-out judgments from the Hebart subset of the Psych-101 dataset, adhering to the strict key-only evaluation standards set by the Centaur framework [1]. Our approach tests semantic-label augmentation, embedding-based similarity measures, and preference learning informed by reasoning to explore how semantic and reasoning signals shape human-aligned decisions. Since key-only evaluation prohibits using semantic labels and explicit reasoning, we focus on analyzing similarity structure, discovering that semantic distance is a strong predictor of human choices in clear-cut cases. Based on this, we indirectly introduce reasoning signals via Direct Preference Optimization (DPO), which yields modest but consistent improvements, even though the small model's reasoning abilities are limited and inconsistent. Overall, these findings suggest that small LLMs, when guided by semantic cues and preference learning, can capture important features of human decision-making. This opens up new possibilities for cognitively-informed alignment within constrained evaluation environments.

1 Introduction

Large language models (LLMs) have recently attracted attention not just for their capabilities in language understanding, but also as computational models for human cognition. The Centaur framework[1], for instance, showed that a 70-billion-parameter Llama model could mirror human behavioral patterns in the Psych-101 dataset, analyzing over 10 million human responses across 160 cognitive tasks. Yet, much of Centaur's performance seems driven by statistical pattern recognition instead of true conceptual reasoning, and the immense computational resources needed for such large models pose challenges for wider use.

In this study, we ask whether a much smaller model can approximate human decision-making in the odd-one-out task, which involves judging relational similarity between objects. Since this task is fundamentally semantic and rooted in human cognition, it provides a strong basis for testing whether semantic cues or reasoning-based signals can improve human-like alignment, especially in constrained evaluation environments. Our central research question is therefore:

> *Can a small LLM learn human-like odd-one-out decision patterns when supported by semantic similarity cues and reasoning-based preference signals?*

To address this question, we investigate five modeling strategies (M1–M5) that incorporate semantic augmentation, similarity cues, or reasoning-based preference signals. Only the key-only baseline (M1) and the preference-aligned model (M5) can be directly tested under Centaur's key-token constraints. The other approaches (M2–M4), while not directly evaluable, offer conceptual perspectives on how semantic and reasoning factors might shape human-aligned decision-making.

2 Related Work

2.1 Psych-101 and the Centaur Framework

The Psych-101 dataset compiles extensive human behavioral responses collected from a diverse array of cognitive tasks. The Centaur framework [1] showed that large LLMs are capable of approximating these human choices by applying supervised fine-tuning using key-only labels (A/B/C). Although this evaluation setup promotes consistency, it limits the integration of semantic or reasoning-based enhancements, as models are required to generate only key tokens (A/B/C) during assessment.

2.2 Odd-One-Out Tasks

The Hebart odd-one-out dataset [2] presents participants with three objects and asks them to choose the one that is least similar to the others. Since this decision relies on relational similarity and category boundaries, the task serves as a valuable benchmark for assessing whether LLMs can mirror human semantic judgments in a controlled setting.

2.3 Parameter-Efficient Fine-Tuning

This study adopts QLoRA [3], a parameter-efficient fine-tuning technique that applies low-rank adapters to quantized model weights. By lowering the computational demands of training, QLoRA enables the fine-tuning of 2–7B parameter models on modest hardware, while maintaining strong empirical results.

3 Theoretical Foundations

This study is based on the assumption that human odd-one-out decisions reflect underlying semantic and reasoning structure. We also assume that elements of this structure can be indirectly found in lexical labels, distributional embeddings, and the explanatory reasoning generated by larger pretrained models. This theoretical viewpoint helps explain why semantic labels, similarity margins, and preference signals derived from reasoning may offer valuable information for a small LLM during training, even if these signals are not directly available at evaluation time.

Lexical Semantics. Object names and their embeddings are assumed to encode coarse semantic information such as attributes and category membership (e.g., animals, tools). Because these representations often form meaningful clusters in embedding space, they provide a conceptual basis for semantic-label augmentation (M2)

and help explain why similarity-based or preference-based methods (M3, M5) may capture aspects of human categorical structure.

Distributional Semantics. Sentence-BERT embeddings [4] place words and phrases in a continuous semantic space that is often assumed to reflect aspects of human similarity judgments. Under this assumption, embedding distances may provide useful signals for margin-based similarity analysis, helping to interpret why trials with clearer semantic separation tend to show more consistent human choices.

Reasoning Signals. For M4, reasoning traces are supplied by a larger external LLM. These explanations typically highlight contrasts among the presented items or articulate relevant category distinctions. Such signals can guide the small model during training, even though the reasoning itself is not used at evaluation time.

Cognitive Semantics. From this perspective, M3 reflects conceptual distance through embedding margins, while M5 learns which distinctions matter through preference comparisons. Both provide partial signals of the conceptual structure behind human odd-one-out judgments.

4 Methodology

4.1 Dataset

We use approximately 10,000 samples from the Hebart odd-one-out dataset. Each trial presents three objects, and participants select the item least similar to the other two. Because the task depends on semantic relations among objects, it offers a suitable setting for studying whether a small LLM can learn patterns that resemble human choices. We sample a subset of trials and convert them into the key-only format required for Centaur-style experimental setup.

4.2 Model Approaches (M1-M5)

We consider five modeling approaches that represent different ways of introducing semantic or reasoning-based augmentation. These approaches are presented to articulate the conceptual landscape of possible extensions, although not all are implemented or trained in practice. Among them, only the key-only baseline (M1) and the DPO model (M5) are fully trained and evaluated under Centaur-style key-only constraints. Below, we describe each approach and its compatibility with the evaluation framework.

M1 - Key-Only Baseline. M1 is a supervised fine-tuning baseline that uses symbolic labels (A/B/C), following the standard Centaur format. This approach is fully compatible with key-only evaluation and serves as the primary reference model for comparison. Because it includes no semantic information, the model relies solely on patterns present in the key-based training data.

M2 - Semantic Label SFT. M2 conceptually replaces symbolic labels with object names to introduce lexical semantic grounding. Because the resulting outputs no longer correspond to discrete key tokens, the key-only NLL metric becomes undefined under Centaur constraints. Therefore, M2 is excluded from final evaluation.

M3 - Semantic + Similarity Augmentation. M3 conceptually incorporates Sentence-BERT similarity scores to provide relational cues among the objects. However, this semantic information is present only during training, while evaluation under Centaur uses strictly key-only prompts. This mismatch between training and test conditions leads us to exclude M3 from the final NLL-based evaluation.

M4 - Reasoning-Augmented Input. M4 conceptually supplements training prompts with short reasoning explanations generated by GPT-4.1-mini. However, reasoning-based augmentation cannot be evaluated under the key-only Centaur framework for three distinct reasons:

- **Train-test mismatch (input reasoning removed at test time):** Reasoning appears in the training prompts but cannot be included in the key-only evaluation format. This mismatch means the model is trained on reasoning-rich inputs but tested on reasoning-free inputs, invalidating NLL comparison.
- **Prompt-format incompatibility (input reasoning kept at test time):** Adding reasoning to the test prompt breaks the strict key-only Centaur format. Since all other models follow this standardized prompt, M4 would not be comparable under the same evaluation metric.
- **Output ambiguity (reasoning generated as output):** Because the training target for M4 includes both reasoning text and the key, the model is not constrained to end its output with a the key token. If the final generated token is not the key, the model's selected label cannot be determined, and the key-level NLL cannot be computed.

Accordingly, M4 is excluded from final evaluation.

M5 - Direct Preference Optimization (DPO). M5 incorporates background reasoning signals while still adhering to the strict key-only format required for Centaur evaluation. Rather than modifying the input prompt, it uses reasoning traces only during training to construct preference data. For each key-only prompt, the base model generates several candidate responses that include a brief internal rationale paired with a selected key. These reasoning–answer chains are used exclusively for preference construction. Claude-sonnet-4.5 evaluates the candidates based on key correctness, clarity of reasoning, and the alignment between the explanation and the chosen item, producing chosen and rejected pairs that guide DPO fine-tuning. Importantly, although reasoning is involved during training, it is completely removed from the inference process. During evaluation, M5 outputs only a single key token and therefore remains fully compatible with the Centaur key-only NLL protocol.

4.3 Training Setup

All models are fine-tuned using QLoRA [3] for efficient training of the Gemma-2B-Instruct architecture. Unless otherwise noted, all variants share the same optimization and hyperparameter settings.

- Reasoning signal generation is performed using GPT-4.1-mini.
  - M4 (Reasoning-SFT) was implemented as a prototype, but it is not included in the final evaluation due to incompatibility with key-only inference and unstable test-time behavior.
  - M5 (Reasoning-DPO) uses reasoning traces only for constructing preference pairs. Claude is prompted to select

the more coherent, consistent, and accurate reasoning-and-answer chain. Reasoning is removed at inference time to maintain key-only evaluation compatibility.

- Exploratory models M2-M4 are used solely for conceptual and diagnostic analysis because they do not satisfy key-only inference constraints.
  - M2 (Semantic) introduces semantic grounding but cannot be evaluated under key-only conditions.
  - M3 (Semantic + Similarity) incorporates Sentence-BERT similarity features. Its accuracy trends indicate potential utility, but it is not included in the final evaluation.
- M1 and M5 are the only models compatible with key-only inference and thus the only candidates for NLL-style evaluation. However, they are not directly comparable because M1 is trained through supervised fine-tuning while M5 applies DPO to the unfine-tuned base model.

4.4 Evaluation Setup

Evaluation follows the Centaur key-only framework, in which the model must output exactly one symbolic label (A/B/C). The primary metric is the negative log-likelihood (NLL) computed on the final token corresponding to the predicted key:

$$\text{NLL} = -\log P_\theta(y_{\text{key}} \mid x_{\text{prompt}}).$$

This metric is valid only when (i) the model outputs a single key token, and (ii) the evaluation prompt strictly matches the Centaur key-only format.

These constraints exclude model variants whose training or inference behavior does not preserve the key-only format. Any approach that introduces additional semantic inputs, similarity-based features, or reasoning-augmented outputs cannot be evaluated under Centaur's NLL evaluation.

Although M1 is trained to produce key-only predictions, it is not included in the quantitative NLL evaluation, only the base Gemma-2B model and the DPO-trained model are evaluated under the Centaur protocol.

5 Key Challenges and Conceptual Fixes

This section explains why the semantic and reasoning augmentations introduced in Section 4 (M2–M4) cannot be evaluated under the Centaur key-only framework, and why an indirect, preference-based solution (M5) becomes necessary. The key issue is that the evaluation metric (NLL) imposes strict requirements on the allowable output format. Methods that incorporate semantic or reasoning information during training cannot make use of these signals at inference time without breaking the key-only constraints, resulting in a train–test mismatch. These limitations motivate a training-time mechanism that incorporates semantic and reasoning signals indirectly while still preserving the key-only evaluation design.

5.1 The Constraint of Key-Only Evaluation

A central challenge in this project arises from the strict evaluation design inherited from Centaur, which computes negative log-likelihood (NLL) over a single symbolic key:

$$\text{NLL} = -\log P_\theta(y_{\text{key}} \mid x_{\text{prompt}}), \tag{1}$$

where x_{prompt} follows the key-only format and $y_{\text{key}} \in \{A, B, C\}$. Because evaluation depends exclusively on the predicted key token, any augmentation that changes the expected output distribution or alters the prompt format invalidates NLL as a comparable metric. This restriction produces three major implications:

- Reasoning is not measurable under NLL. The NLL metric evaluates only the conditional probability of the final key token. Formally, Eq. (1) computes $-\log P_\theta(y_{\text{key}} \mid x_{\text{prompt}})$, where x_{prompt} contains no semantic or reasoning cues. If the model internally generates a reasoning sequence r before selecting a key, the evaluation marginalizes over r:

  $$P_\theta(y_{\text{key}} \mid x_{\text{prompt}}) = \sum_r P_\theta(r, y_{\text{key}} \mid x_{\text{prompt}}),$$

  so differences in reasoning quality have no effect on the score.
  Separately, any attempt to add semantic labels or reasoning text to the prompt changes x_{prompt} itself. Such augmentation violates the key-only requirements of the Centaur framework and makes the resulting model incompatible with NLL-based evaluation.
- Semantic signals cannot appear at inference time. Any semantic or reasoning information provided during training must be absent from the test-time prompt to preserve the key-only format, creating a fundamental train-test mismatch. This mismatch makes it impossible to determine whether gains during training reflect genuine improvements in conceptual alignment or artifacts of unavailable cues.
- Changing the output space breaks the metric. NLL in Eq. (1) is defined only over the key vocabulary $y_{\text{key}} \in \{A, B, C\}$. Methods such as M2, which replace symbolic keys with lexical object names y_{word}, alter the output token space:

  $$y_{\text{key}} \;\rightarrow\; y_{\text{word}},$$

  removing the target token from the evaluation domain. In this case, $P_\theta(y_{\text{key}} \mid x_{\text{prompt}})$ is no longer defined, making NLL impossible to compute or compare.

Thus, although semantic and reasoning augmentations enrich the training process, they cannot be used directly during evaluation without violating the key-only constraints of the Centaur framework.

5.2 Limitations of Semantic Augmentation

In summary, the semantic and reasoning augmentations used in M2–M4 cannot be applied during evaluation because the Centaur evaluation requires key-only prompts and single-token outputs. Any information added during training—semantic labels, similarity features, or reasoning text—must be absent at test time, which creates a train–test mismatch and makes these approaches incompatible with the NLL metric. This limitation motivates an indirect training method (M5) that incorporates such signals without altering the key-only evaluation format.

5.3 Limitations of Direct Reasoning Augmentation

M4 incorporates GPT-generated reasoning traces into the training prompts, but these cues cannot appear at inference time under the key-only evaluation protocol. Because the metric requires the prediction of a single key token, reasoning-augmented formats used during training cannot be carried over to evaluation. Thus, M4 is useful as a training signal but cannot be evaluated directly under the Centaur key-only metric.

5.4 Motivation for DPO as an Indirect Solution

Since semantic cues (M2, M3) and direct reasoning augmentation (M4) cannot be incorporated into the evaluation pipeline without violating the key-only NLL constraints, an indirect solution is required. Direct Preference Optimization (DPO) enables this separation: the training process can use comparisons of full reasoning–and–answer chains, while these chains do not appear during inference. In other words:

- training may incorporate semantic and reasoning cues,
- evaluation enforces a strict key-only output format.

This makes M5 a suitable variant that uses reasoning-related training signals without violating the Centaur evaluation protocol.

5.5 Preference Data Construction for DPO

Because the Psych-101 dataset contains no preference labels, we construct synthetic preferences in two stages:

(1) The base model generates multiple reasoning–and–key responses for each prompt.
(2) An external LLM (Claude) compares pairs of responses and determines which one is preferred, based on correctness, logical soundness, and consistency.

The resulting preferred/rejected pairs serve as training signals for DPO. Consistent with RLHF practices, the effectiveness of these signals depends on the coherence of the generated reasoning.

5.6 Training Objective Structure

In our experiments, SFT and DPO are trained as separate models rather than sequential stages. The supervised key-only model (M1) is trained with standard supervised fine-tuning to predict the key label, but it is not used as the starting point for DPO. Instead, DPO is applied independently to the unfine-tuned Gemma-2B base model using the synthetic preference pairs. This separation ensures that key-only output behavior is preserved during evaluation and that the SFT and DPO objectives do not interact at the gradient level.

5.7 Evaluation Metrics and Constraints

Although NLL cannot capture semantic understanding or human-like reasoning, it is the only metric compatible with the Centaur key-only evaluation setting. Therefore, NLL is used exclusively for final quantitative comparison. Other matrics remain outside the scope of this evaluation and are left for future work.

5.8 Summary of Conceptual Fixes

Overall, these challenges explain why:

- M2-M4 cannot be evaluated under the key-only constraints,
- their semantic and reasoning information remains valuable,
- motivating M5, where reasoning signals are incorporated indirectly through DPO while maintaining key-only inference.

DPO thus serves as a principled mechanism to integrate reasoning-based preferences without violating Centaur's evaluation

6 Results

6.1 Overview

This section presents empirical findings from the five modeling approaches (M1–M5). Among these, only the raw base model and the DPO-aligned model (M5) are directly evaluated using negative log-likelihood (NLL), since they follow the key-only evaluation protocol. The supervised key-only model (M1) is trained via supervised fine-tuning to predict the key label, but it is not included in the NLL comparison because DPO is applied only to the unfine-tuned base model.

Models M2-M4 provide conceptual insight into how semantic and reasoning signals might influence odd-one-out decisions. To capture these contributions, we analyze (1) Sentence-BERT-based semantic similarity patterns (M3) (2) preference-based reasoning alignment through DPO (M5).

6.2 Semantic Similarity Analysis (M3)

Although M3 cannot be evaluated directly under the NLL metric, the semantic similarity experiments reveal notable behavioral patterns that shed light on how humans make odd-one-out decisions and how embeddings approximate this behavior.

6.2.1 Embedding Space Visualization. Figure 1 displays a t-SNE projection of sample object embeddings. The distribution shows no clear global clusters, suggesting that global category structure does not drive the task. Instead, *local relational distances* among the three presented items appear to determine the odd-one-out choice.

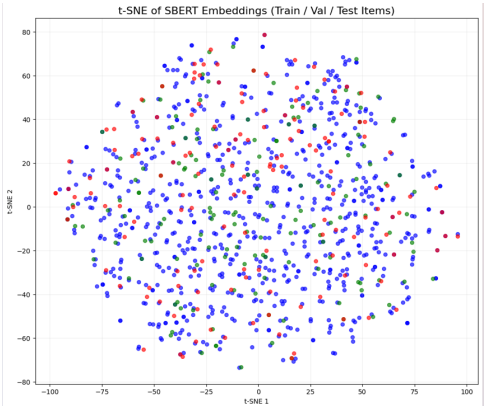

Figure 1: t-SNE visualization of SBERT embeddings. The distribution suggests odd-one-out decisions rely on local semantic relationships rather than global clusters.

6.2.2 Sentence-BERT Confusion Matrix. SBERT distance-based labels achieve an accuracy of 0.43 against human choices, above the random baseline of 0.33. This suggests that SBERT embeddings contain weak but consistent decision cues.

6.2.3 Margin-Based Accuracy. We examine how semantic separation affects agreement between SBERT-derived labels and human choices. For each trial:

$$\text{margin} = d_{\text{max}} - d_{\text{next}}.$$

Higher margins indicate one object is more semantically distant, and such trials show higher agreement with human labels.

Threshold-Based Accuracy. As the similarity margin increases, SBERT-derived choices align more closely with human labels. High-margin trials show much higher accuracy.

Margin Threshold	Size (n)	Accuracy
≥ 0.02	3953	0.473
≥ 0.05	665	0.617
≥ 0.08	130	0.731
≥ 0.10	52	0.712

Table 1: Accuracy grouped by semantic margin threshold.

Top-K High-Margin Accuracy. These high-margin subsets show substantially higher alignment with human labels compared to the overall average.

Subset (Top-K)	Size (n)	Accuracy
Top 20%	2000	0.514
Top 10%	1000	0.582
Top 5%	500	0.654
Top 2%	200	0.730
Top 1%	100	0.740

Table 2: Accuracy on highest-margin subsets.

Both analyses converge:

> *When semantic distance is large, humans and embeddings choose the same odd-one-out item with high consistency.*

Implication for Future Work. Since high-margin cases show strong human–model alignment, future systems may benefit from:

- margin-weighted loss functions-increasing weight on high-margin examples,
- confidence-aware training-downweighting ambiguous (low-margin) cases.

Such methods could help small LLMs form more stable semantic decision boundaries.

6.3 Preference-Based Reasoning Alignment (M5)

M5 uses Direct Preference Optimization (DPO) to incorporate reasoning information *indirectly* while keeping evaluation strictly key-only. Reasoning traces are generated during training, compared by an external LLM judge (Claude-sonnet-4.5), and used to create preference pairs. However, at inference time, the model outputs only a single key token, making it fully compatible with Centaur-style NLL evaluation.

6.3.1 DPO Pipeline. Figure 2 summarizes the end-to-end pipeline.

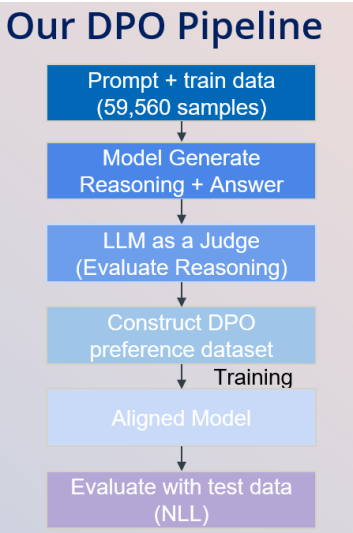

Figure 2: DPO training pipeline: reasoning + answer generation, preference judgment, preference dataset construction, DPO fine-tuning, and key-only evaluation.

6.3.2 DPO Evaluation Results. Although M1 satisfies the key-only format, it is excluded from quantitative evaluation. We use the unfine-tuned Gemma-2B model as the baseline for NLL comparison on the same Psych-101 test set.

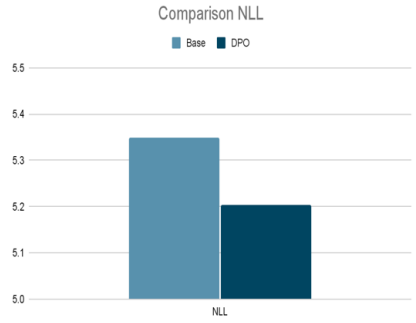

Figure 3: Comparison of NLL for baseline (Gemma-2b) and DPO-enhanced model (M5).

- Baseline (Gemma-2b) NLL: **5.349**
- DPO-enhanced (M5) NLL: **5.203**

The reduction in NLL shows that DPO improves key-only calibration despite the absence of reasoning at test time. The improvement is modest, likely due to noisy small-model reasoning and limited preference data, indicating room for more robust preference supervision in future work.

6.4 Summary of Findings

- Semantic similarity signals (M3) strongly correlate with human decisions in high-margin cases, confirming that embeddings encode meaningful conceptual structure.
- Reasoning cannot be used directly under the Centaur evaluation regime, but preference-based reasoning signals (M5) successfully improve key-only performance.
- DPO provides a principled mechanism for incorporating reasoning benefits while maintaining evaluation compatibility.
- The limited improvement suggests that better initial reasoning and richer preference datasets are crucial for unlocking the full potential of preference-based alignment.

7 Discussion

Key observations are summarized below:

- High-margin trials show strong alignment between Sentence-BERT similarity and human choices, indicating that semantic distance provides informative signals that may be useful for future modeling.
- Direct reasoning cannot be used under key-only scoring. DPO offers an indirect alternative: it incorporates reasoning during training while preserving key-only outputs, yielding modest but consistent improvements.
- Semantic and reasoning cues are informative yet restricted by key-only evaluation. Preference-based methods provide a practical way to exploit these signals within Centaur's constraints.

8 Limitations

- The key-only evaluation framework prevents the use of semantic labels or reasoning text during training or inference, limiting our ability to assess or integrate these signals directly.
- Due to computational constraints, experiments were conducted on a reduced subset of the Psych-101 dataset, increasing uncertainty in performance estimates.
- DPO supervision relies on synthetic preferences generated from small model reasoning traces, which provide weak and inconsistent signals and limit the strength of preference based alignment.

9 Future Work

- Margin-aware training. Semantic margins that align with human judgments could be used to weight training samples, incorporating semantic reliability into the learning objective.
- Expanded embedding exploration. Additional embedding models (e.g., CLIP, LLM-based embeddings, or domain-specific semantic spaces) could be examined to determine whether alternative representations yield stronger or more robust semantic–behavioral alignment.
- Stronger preference signals. Because DPO relies on reasoning-based comparisons, improving the quality and consistency of reasoning traces may lead to more effective preference alignment.
- Broader evaluation and data coverage. Using larger portions of the Psych-101 dataset or extending to additional decision tasks may help assess the generality of these findings.
- After addressing resource and time constraints, I plan to build DPO-augmented SFT models on a shared dataset and examine whether adding DPO offers any advantage over SFT alone, as we hypothesize.

10 Conclusion

This study examined whether small LLMs can approximate human decision patterns in odd-one-out tasks by incorporating semantic similarity and reasoning-based supervision. Although semantic- and reasoning-augmented SFT models (M2–M4) were incompatible with Centaur's key-only evaluation, their analyses revealed that semantic margins align strongly with human choices and that reasoning traces contain useful supervisory signals.

Direct Preference Optimization (M5) offered a way to use these signals indirectly while preserving key-only outputs. M5 produced modest but consistent improvements in NLL, indicating that preference-based alignment can enhance decision calibration even under evaluation constraints.

Overall, the results show that small LLMs can benefit from semantic cues and preference-guided reasoning when modeling human-like decisions.

11 Code Availability

All code used in this study is available at: GitHub Repository

12 Data Availability

- Psych-101 Dataset (Train): Hugging Face Datasets
- Psych-101 Dataset (Test): Hugging Face Datasets

All code used in this study is available at: GitHub Repository

References

[1] Michael Binz, Eric Schulz, and David Krueger. Centaur: Learning human decision patterns with large language models. *Nature*, 2025.

[2] Martin N. Hebart, Chaitanya Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. *Nature Human Behaviour*, 4(11):1173–1185, 2020.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized large language models. *Advances in Neural Information Processing Systems*, 2023.

[4] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, 2019.

A Appendix

Q1. How did you gather preference signals for DPO? Does the original Psych-101 dataset contain ground-truth preference pairs, or were these constructed artificially (e.g., via Claude)? How valid is this preference supervision?

Since the Psych-101 dataset does not contain ground-truth preference pairs, all DPO preference signals were constructed artificially. For each key-only prompt, the base model generated multiple candidate responses that included a brief reasoning trace and a selected key. These candidates were then evaluated by an external judge (Claude-sonnet-4.5), which selected the preferred response based on key accuracy, reasoning coherence, and the consistency between the explanation and the chosen item. The chosen–rejected pairs produced through this comparison served as the preference data for DPO fine-tuning. To evaluate the reliability of these synthetic labels, a sample of the preference pairs was manually reviewed. Human inspection confirmed that the external judge's decisions aligned with the intended evaluation criteria, supporting the use of the same rubric to automatically generate the remaining preference pairs. Although the reasoning traces produced by the small model were sometimes weak or inconsistent, the preference comparisons still captured meaningful differences in response quality, making them appropriate for DPO training.

Q2. Do you optimize both NLL (for next-token accuracy) and DPO (for preference alignment)? If so, how do the gradients interact, and what objective takes priority during training?

We do not optimize NLL and DPO jointly. In our experiments, SFT and DPO were trained as separate models rather than sequential stages. The SFT model was trained with supervised fine-tuning to predict the key-only label, reaching a token-level training loss of 0.83. This training loss reflects the SFT objective and is not comparable to key-only NLL, which we report separately.

Due to resource constraints, we did not apply DPO on top of the SFT checkpoint. Instead, DPO was applied directly to the untrained Gemma-2b base model, where it reduced key-only NLL relative to the baseline despite the weak reasoning traces produced by the smaller model. Because the two objectives were optimized independently, their gradients do not interact.

Based on the improvement observed in the DPO-trained base model, we expect that combining SFT with DPO would further reduce key-only NLL, which we identify as a promising direction for future work.

Q3. Since NLL only measures probability of the correct key token-not semantic reasoning, not human alignment, why is it considered an appropriate evaluation metric for human-like decision modeling? What alternative metrics did you consider?

In this project, NLL was used not because it captures semantic reasoning or human-like alignment, but because it is the only metric compatible with the Centaur key-only evaluation protocol. Under this setup, the model must output exactly one symbolic label (A, B, or C), and evaluation is based solely on the log-probability of that final token. As a result, NLL serves as the standard metric for fair comparison with prior work, even though it does not reflect deeper cognitive processes.

We acknowledge that NLL alone cannot fully characterize human-like decision behavior. For this reason, we examined several complementary metrics. First, we analyzed semantic similarity patterns using Sentence-BERT and found that human–model agreement increases in high-margin trials, providing insight into the model's grasp of conceptual structure. We also considered behavior-based measures such as embedding-derived prediction accuracy and per-trial agreement rates. Looking forward, alternative metrics such as rank correlations between model logits and human choice distributions, or semantic alignment measures that evaluate how closely model representations mirror human similarity judgments, could offer more cognitively meaningful evaluation. Since these metrics fall outside the Centaur protocol, we report them only as supplementary analyses rather than primary evaluation scores.

Q4. The project relies heavily on LLM prompting and DPO. Where exactly does NLP theory enter e.g., semantics, pragmatics, lexical modeling, distributional structure? Could you clarify the NLP foundations?

Although the project makes extensive use of LLM prompting and DPO, its foundation rests on several core areas of NLP theory.

First, lexical semantics plays a central role. Odd-one-out judgments reflect category structure, attribute similarity, and relational meaning, which aligns with how lexical items encode semantic features. This provides a theoretical basis for using object labels and semantic augmentation.

Second, the project relies on distributional semantics. Our analyses using Sentence-BERT assume that words appearing in similar contexts occupy nearby regions in the embedding space. The margin-based similarity experiments test this assumption directly and show that human choices correlate with geometric structure in the embedding space, particularly in high-margin cases.

Third, the reasoning components are closely related to pragmatic theory. Preference comparisons evaluate whether a model's explanation is coherent, relevant, and consistent with the chosen option. These criteria reflect principles from classical pragmatics and modern approaches to natural language inference.

Finally, the key-only supervised setup reflects foundational ideas in lexical modeling and probabilistic language modeling, where

next-token likelihood is used to assess how well a model represents symbolic distinctions.

Although we briefly explored retrieval-augmented generation (RAG) as a potential extension, we ultimately excluded it because external retrieved information cannot be incorporated under the Centaur key-only protocol. RAG does not align well with the task's theoretical assumptions, which center on lexical and distributional semantics rather than external knowledge retrieval.

We also considered analyzing attention patterns using BertViz to investigate which lexical or semantic cues the model focused on during decision-making. After shifting our emphasis toward DPO-based preference alignment, we were unable to include this analysis due to time constraints. Attention visualization remains a promising direction for future work, as it may offer additional insight into how small LLMs allocate attention across the three objects in odd-one-out tasks.

Together, these elements show that the project is grounded in established areas of NLP theory. Although implementation involves prompting and alignment techniques, the work is directly involved in lexical semantics, distributional structure, pragmatic reasoning, and token-level probability modeling.