# LEVERAGING RAG FOR TRAINING-FREE ALIGNMENT OF LLMS

**John T. Halloran**
Leidos
halloranjt@leidos.com

## ABSTRACT

We introduce *R*etrieval *A*ugmented *G*eneration for *Pref*erence alignment (RAG-Pref), a training-free alignment algorithm compatible with existing off-the-shelf packages. By conditioning on preferred and dispreferred samples during inference, RAG-Pref utilizes additional *contrastive information* compared to standard RAG. For agentic safety alignment across five widely used models, we show that while state-of-the-art offline (training-based) and online preference alignment algorithms struggle to improve refusal guardrails against adversarial attacks, RAG-Pref effectively improves refusal performance. Furthermore, in stark contrast to other online alignment algorithms, RAG-Pref drastically improves performance on general human-preference alignment tasks without substantially increasing computational requirements.

## 1 INTRODUCTION

Alignment has become a critical step towards ensuring the responses of large language models (LLMs) align with general human preferences (Zheng et al., 2023). Currently, alignment algorithms are dominated by reinforcement learning-based schemes such as RLHF (Ouyang et al., 2022) and the computationally efficient direct preference optimization (DPO) (Rafailov et al., 2023), wherein models are post-trained over pairs of preferred and dispreferred responses. Such *alignment-tuning* has proven pivotal towards producing LLM assistants whose responses are both accurate and helpful (Ouyang et al., 2022; Zheng et al., 2023; Achiam et al., 2023).

Given LLMs' susceptibility to adversarial attacks (Mehrotra et al., 2024; Chao et al., 2025), significant work has focused on safety alignment-tuning (SAT) to enable *refusal guardrails* (Bai et al., 2022; Ji et al., 2023; Dai et al., 2024; Kim et al., 2025). However, with the advent of the *model context protocol* (MCP) (Anthropic, 2025b), new security risks have emerged (Kumar et al., 2025; Radosevich & Halloran, 2025). In particular, new attacks which induce malicious tool use yet lack standard refusal trigger phrases, called *falsely benign attacks* (FBAs), have proven highly effective against MCP-enabled LLMs.

Collecting FBAs and truly benign (TB) samples, we show that SOTA SAT algorithms display limited ability to enable FBA refusal guardrails—across all models, DPO and SafeDPO improve baseline refusal rates by only average factors of 1.4 and 1.6, respectively. Alarmingly, no safety-tuned model achieves an FBA refusal rate greater than 48%. To address this, we introduce *R*etrieval *A*ugmented *G*eneration for *Pref*erence alignment (RAG-Pref), which utilizes RAG to *contrastively condition* on both preferred and dispreferred examples during inference. Compared to standard RAG, RAG-Pref is guaranteed to further reduce expected uncertainty during inference by a nonnegative amount, referred to as the *contrastive information*.

RAG-Pref is online (training-free), easily implementable using off-the-shelf packages, and significantly improves refusal guardrails. When combined with DPO and SafeDPO aligned models, RAG-Pref increases baseline refusal rates by average 3.5- and 3.9-fold improvements, respectively.

We show that RAG-Pref similarly provides drastic performance improvements for general human-preference alignment tasks, leading to an average 34.9% and 3.3% increase in AlpacaEval 2 and MT-Bench performance, respectively, across SOTA alignment-tuned models. Contrasted with the computational requirements of other offline and online alignment algorithms, RAG-Pref requires

three orders of magnitude less preparation time than DPO and only a 20% inference slowdown (versus 372% for other recent online methods (Zhu et al., 2025)).

## 2 Background and Related Work

**Alignment and Safety.** LLM alignment near ubiquitously consists of fine-tuning given queries with accompanying preferred and dispreferred response pairs. Initial approaches like RLHF (Ouyang et al., 2022) trained reward models then used RL fine-tuning. DPO (Rafailov et al., 2023) reparameterized the RL objective, allowing direct learning with a simple closed-form objective. DPO and its variants (Melnyk et al., 2024; D'Oosterlinck et al., 2025; Jung et al., 2024; Ji et al., 2024) have demonstrated strong performance on benchmarks like AlpacaEval 2 (Dubois et al., 2024) and MT-Bench (Zheng et al., 2023).

Safety-focused work has extended alignment to harmful behaviors (Hartvigsen et al., 2022; Mazeika et al., 2024), with SafeDPO (Kim et al., 2025) optimizing for safety via additional loss offsets.

For online training-free alignment, (Zhu et al., 2025) introduced On-the-fly Preference Alignment via Principle-Guided Decoding (OPAD). OPAD calculates a similar reward function as used in DPO to adjust the per-token conditional distribution during decoding and was shown to improve performance on general preference alignment tasks relative to previous online methods (Gao et al., 2024).

**Agentic Attacks.** Jailbreaks (Zou et al., 2023; Chao et al., 2025) and prompt injection attacks (Perez & Ribeiro, 2022; Greshake et al., 2023) have been extensively studied. However, frontier LLM safety training has grown to include standard attacks (Mazeika et al., 2024; Chao et al., 2024), expanding refusal guardrails. Recently, the MCP has seen massive adoption (Anthropic, 2025a), enabling seamless integration between AI agents and applications. However, Radosevich & Halloran (2025) demonstrated that MCP-enabled agents are susceptible to LLM attacks which explicitly lack refusal guardrail triggers, i.e., FBAs. The success of FBAs is attributed to the shift in attack goals from LLMs–which focus on unsafe text generation, the attacks of which contain related harmful phrases or suspicious text–to MCP-enabled LLMs–which focus on the malicious execution of tools, the attacks of which need not contain harmful or suspicious text found in standard SAT data.

## 3 RAG-Pref: Training-Free Alignment

Let $\mathcal{D} = \{(x^i, y_w^i, y_l^i)\}$ be preference data where $y_w^i \succ y_l^i | x$. DPO trains models by optimizing $\mathcal{L}(r_\phi) = -\mathbb{E}[\log \sigma(\beta r_\phi(x, y_w) - \beta r_\phi(x, y_l))]$ where $r_\phi(x, y) = \log \frac{\pi_\phi(y|x)}{\pi_{\text{ref}}(y|x)}$. OPAD (Zhu et al., 2025) adjusts per-token distributions using DPO's reward function $r_\phi(x, y)$ during decoding, thus achieving alignment while avoiding training. However, OPAD requires invasive changes to generation as well as incurring heavy computational overhead, requiring storage of two additional token distributions during decoding.

RAG-Pref performs online alignment without invasive generation changes and heavy additional computational requirements. Let $e(\cdot)$ be an embedding function and $d(\cdot, \cdot)$ a distance metric. We construct vector databases $\mathcal{D}_e^w$ and $\mathcal{D}_e^l$ from preferred and dispreferred responses, respectively. The RAG-Pref algorithm is listed in Algorithm 1. Compared to alternative alignment algorithms: (1) DPO enforces preference alignment over response *pairs* $(y_w, y_l)$, while RAG-Pref enables alignment over *sets* $\mathcal{Z}^w$ and $\mathcal{Z}^l$; (2) No invasive changes to generation are required, ensuring compatibility with widely-used models and off-the-shelf packages.

**Algorithm 1** RAG-Pref for online alignment.

**Input:** Query $x$, $\mathcal{D}_e^w$, $\mathcal{D}_e^l$, retrieval count $k$.
1: Embed $x$: $x' = e(x)$
2: Retrieve top-$k$ preferred responses $\mathcal{Z}^w$ by ranking $d(x', z)$ for $z \in \mathcal{D}_e^w$
3: Retrieve top-$k$ dispreferred responses $\mathcal{Z}^l$ similarly
4: Create instruction $\mathcal{Z}^w \succ \mathcal{Z}^l$ to follow preferred and avoid dispreferred examples
5: **return** $\pi_\theta(y|x, \mathcal{Z}^w \succ \mathcal{Z}^l)$

Table 1: **FBA Refusal Rates**: Bold = highest per model. GEMMA-2-2B-IT incompatible with OPAD.

| | Offline Only | | | + OPAD | | | + RAG-Pref | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Base | DPO | SafeDPO | Base | DPO | SafeDPO | Base | DPO | SafeDPO |
| LLAMA-3.2-1B-INSTRUCT | 0.15 | 0.31 | 0.40 | 0.59 | 0.61 | 0.66 | 0.28 | 0.58 | **0.88** |
| GEMMA-2-2B-IT | 0.32 | 0.45 | 0.47 | – | – | – | 0.63 | 0.74 | **0.75** |
| LLAMA-3.1-8B-INSTRUCT | 0.35 | 0.43 | 0.45 | 0.43 | 0.47 | 0.37 | 0.95 | **0.97** | **0.97** |
| DEEPSEEK-R1-DISTILL-LLAMA-8B | 0.14 | 0.15 | 0.13 | 0.44 | 0.47 | 0.45 | 0.59 | 0.59 | **0.59** |
| DEEPSEEK-R1-DISTILL-QWEN-14B | 0.16 | 0.18 | 0.19 | 0.41 | 0.44 | 0.46 | 0.64 | **0.68** | 0.64 |

**Contrastive Information.** RAG-Pref performs *contrastive conditioning*—conditioning on both positive examples (what to do) and negative examples (what to avoid). Standard RAG lacks this contrastive conditioning. Let $\Delta H_{\text{RAG}} = I(Y; Z^w|X)$ and $\Delta H_{\text{RAG-Pref}} = I(Y; Z^w, Z^l|X)$. We define the *contrastive information* as $\Delta H_{\text{RAG-Pref}} - \Delta H_{\text{RAG}}$.

**Theorem 3.1.** *$\Delta H_{RAG\text{-}Pref} \geq \Delta H_{RAG}$. When dispreferred examples provide non-redundant information compared to preferred examples, $\Delta H_{RAG\text{-}Pref} > \Delta H_{RAG}$.*

The proof is in Appendix A. For safety alignment, attack patterns are often semantically distinct from refusal responses, providing substantial contrastive information. Theorem 3.1 explains why standard RAG can *decrease* safety (An et al., 2025)—it retrieves attack examples without contrastive refusals, causing models to misinterpret these as behaviors to follow rather than refuse.

## 4 METHODS

**FBA Data.** FBAs were obtained by mapping CVE exploits (Mann & Christey, 1999) to MCP tool sequences for 10 file/directory manipulation tools (Table 5). The MITRE corpus was filtered for agentic attacks—malicious code execution, remote access control, and credential theft—yielding $\sim$34k samples. Using GPT-4O, each CVE was: (a) mapped to Linux commands, (b) marked as feasible given MCP tools, (c) mapped to MCP tool call sequences, and (d) converted to friendly malicious requests. This produced 1,150 FBAs. TB samples were generated by prompting CLAUDE 3.7 SONNET for useful examples per tool. The final dataset: 1,035 training FBAs, 1,035 TB samples, 115 test FBAs. We focus on **refusal guardrail evaluation**, not attack success rates.

**Alignment Setup.** RAG-Pref used off-the-shelf libraries (`ChromaDB`, `LangChain`) and `all-MiniLM-L6v2` embeddings with $k = 2$ for safety and $k = 8$ for general alignment. DPO/SafeDPO used QLoRA (Dettmers et al., 2023) with 15 epochs. OPAD used the official implementation, optimized for batched inference. Details in Appendix E.

## 5 EXPERIMENTS AND CONCLUSIONS

**FBA Refusal Guardrails.**We consider five widely used open-source LLMs, varying in parameter count from 1B to 14B. RAG-Pref is compared to offline (DPO and SafeDPO) and online (OPAD) alignment methods. DPO and SafeDPO refusal alignment were performed using FBA and TB preference pairs. RAG-Pref preferred and dispreferred vector databases were formed from the same preference pairs used for offline alignment. In addition to base models, RAG-Pref and OPAD were combined with offline aligned models. Owing to the use of off-the-shelf components, RAG-Pref was compatible with all evaluated models. In stark contrast, OPAD's invasive decoding scheme was incompatible with GEMMA-2-2B-IT.

Refusal rates across all methods are listed in Table 1. Despite the majority of evaluated base models undergoing excessive post-training safety alignment (Grattafiori et al., 2024; Gemma et al., 2024), no base model achieves an FBA refusal rate over 35%. Furthermore, while refusal rates improve given offline alignment, neither DPO nor SafeDPO enable refusal rates beyond 48%. Thus, **SOTA offline alignment methods provide limited refusal guardrails against FBAs**.

RAG-Pref successfully improves guardrails over all base and SAT models, while OPAD fails to improve LLAMA-3.1-8B-INSTRUCT SAFEDPO guardrails. OPAD outperforms RAG-Pref for the

Table 2: **Human-Preference Alignment**: AlpacaEval 2 and MT-Bench. Bold = top online method. Reported metrics are win rate (WR), length-controlled win rates (LC), and single-answer grading (SAG).

| | | Offline Alignment | | | | |
|---|---|---|---|---|---|---|
| Metric | Online | SFT | DPO | PPO | SimPO | RTO |
| AlpacaEval 2 (WR) | RAG-Pref | **10.87** | **19.50** | **19.07** | **18.14** | **34.29** |
| | RAG | 10.37 | 16.83 | 17.76 | 14.66 | 32.67 |
| | OPAD | 0.99 | 5.59 | 7.95 | 7.95 | 9.62 |
| | – | 9.44 | 12.86 | 14.66 | 10.56 | 31.30 |
| AlpacaEval 2 (LC) | RAG-Pref | **14.48** | **24.82** | **28.59** | **23.18** | **37.45** |
| | RAG | 13.26 | 21.81 | 27.78 | 19.66 | 36.56 |
| | OPAD | 2.26 | 7.61 | 10.16 | 8.41 | 10.69 |
| | – | 14.17 | 14.46 | 18.30 | 16.85 | 36.17 |
| MT-Bench (SAG) | RAG-Pref | **6.01** | **6.19** | **6.49** | **5.84** | **6.83** |
| | RAG | 5.97 | 6.12 | 6.45 | 5.74 | 6.75 |
| | OPAD | 3.58 | 4.05 | 5.14 | 3.16 | 4.58 |
| | – | 5.74 | 6.00 | 6.22 | 5.84 | 6.54 |

base and DPO-aligned LLAMA-3.2-1B-INSTRUCT models. However, RAG-Pref greatly outperforms OPAD all other model configurations, providing an average 50% more refusal performance across the twelve models. Furthermore, for the models OPAD was unable to align, RAG-Pref improved refusal guardrails by an average 74%. Thus, **RAG-Pref drastically outperforms other online methods for the refusal of FBAs, while significantly improving the refusal guardrails of SOTA offline alignment methods**.

**Computational Advantages.** RAG-Pref preprocessing is **7,824× faster** than DPO training. At inference, RAG-Pref is **3× faster than OPAD** and requires **4.2× less memory**. Full results in Table 3 (Appendix C).

**General Human-Preference Alignment.** Table 2 shows AlpacaEval 2 and MT-Bench results for LLAMA-3-8B post-trained using SFT then SOTA offline alignment alignment (DPO, PPO, SimPO, and RTO). RAG-Pref outperforms all online methods across all tasks, improving over baselines by 24.4%, over RAG by 7.3%, and over OPAD by 228.4%. **RAG-Pref is the only online method to consistently improve performance across all tasks and offline algorithms**; OPAD fails to improve baseline performance across all configurations, while RAG fails to improve SFT for win rate and SimPO for MT-Bench. Contrastive information accounts for ∼30% of RAG-Pref's mutual information (Table 4, Appendix D).

### 5.1 CONCLUSIONS

We introduced RAG-Pref, a **training-free, test-time alignment algorithm** implementable with off-the-shelf RAG components. We showed that existing SOTA SAT algorithms are limited in their ability to strengthen refusal guardrails against agentic FBAs. In stark contrast, RAG-Pref drastically improves safety guardrails compared to other online alignment algorithms and offline alignment alone. Furthermore, RAG-Pref is both computationally efficient and easily compatible with widely-used, off-the-shelf packages; RAG-Pref offers 7,824× faster preparation than DPO, while offering 3× faster (with 4.22× less memory) inference than online-alignment competitor OPAD.

Beyond agentic safety, we demonstrated that RAG-Pref similarly improves performance for general human-preference alignment tasks. Across SOTA alignment-tuned models, RAG-Pref leads to consistent improvements on both AlpacaEval 2 and MT-Bench benchmarks, with average increases of 24.4%, 7.3%, and 228.4% over baseline models, standard RAG, and alternative offline alignment method OPAD. Theoretically, we showed that contrastive conditioning provided by RAG-Pref guarantees a reduction of uncertainty during inference, explaining recent work showing RAG degrades safety alignment. Future work includes efficient hyperparameter optimization and the exploration of alternative RAG architectures (Edge et al., 2025; Chan et al., 2025).

**Acknowledgments.** We thank Leidos for funding this research through the Office of Technology. Approved for public release **25-LEIDOS-0521-29630**.

## REFERENCES

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

Bang An, Shiyue Zhang, and Mark Dredze. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5444–5474, 2025.

Anthropic. Donating the model context protocol and establishing the agentic ai foundation. https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation, December 2025a. Accessed: 2026-01-26.

Anthropic. *Introducing the Model Context Protocol*. "https://www.anthropic.com/news/model-context-protocol", 2025b. "Accessed: 2025-02-12".

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Brian J Chan, Chao-Ting Chen, Jui-Hung Cheng, and Hen-Hsen Huang. Don't do rag: When cache-augmented generation is all you need for knowledge tasks. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 893–897, 2025.

Patrick Chao, Edoardo Debenedetti, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TyFrPOKYXw.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *Transactions of the Association for Computational Linguistics*, 13:442–460, 2025.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, et al. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback. In *International Conference on Machine Learning*, pp. 14702–14722. PMLR, 2024.

Team Gemma, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pp. 79–90, 2023.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Haozhe Ji, Cheng Lu, Yilin Niu, Pei Ke, Hongning Wang, Jun Zhu, Jie Tang, and Minlie Huang. Towards efficient exact optimization of language model alignment. *arXiv preprint arXiv:2402.00856*, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24678–24704. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4dbb61cb68671edc4ca3712d70083b9f-Paper-Datasets_and_Benchmarks.pdf.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024.

Geon-Hyeong Kim, Youngsoo Jang, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, and Moontae Lee. Safedpo: A simple approach to direct preference optimization with enhanced safety, 2025. URL https://arxiv.org/abs/2505.20065.

Sonu Kumar, Anubhav Girdhar, Ritesh Patil, and Divyansh Tripathi. Mcp guardian: A security-first layer for safeguarding mcp-based ai system. *arXiv preprint arXiv:2504.12757*, 2025.

David E Mann and Steven M Christey. Towards a common enumeration of vulnerabilities. In *2nd Workshop on Research with Security Vulnerability Databases, Purdue University, West Lafayette, Indiana*, pp. 9, 1999.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, pp. 35181–35224. PMLR, 2024.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105, 2024.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jarret Ross. Distributional preference alignment of llms via optimal transport. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

ProtectAI. *Model Card for distilroberta-base-rejection-v1*. "`https://huggingface.co/protectai/distilroberta-base-rejection-v1`", 2025. "Accessed: 2025-05-15".

Brandon Radosevich and John Halloran. Mcp safety audit: Llms with the model context protocol allow major security exploits. *arXiv preprint arXiv:2504.03767*, 2025.

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

Philipp Schmid. *How to use Anthropic MCP Server with open LLMs, OpenAI or Google Gemini*. "`https://github.com/philschmid/mcp-openai-gemini-llama-example`", 2025. "Accessed: 2025-04-28".

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Mingye Zhu, Yi Liu, Lei Zhang, Junbo Guo, and Zhendong Mao. On-the-fly preference alignment via principle-guided decoding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

## A Proof of Theorem 3.1

**Theorem A.1.** *$\Delta H_{\text{RAG-Pref}} \geq \Delta H_{\text{RAG}}$. When dispreferred examples provide non-redundant information compared to preferred examples, $\Delta H_{\text{RAG-Pref}} > \Delta H_{\text{RAG}}$.*

*Proof.* $\Delta H_{\text{RAG-Pref}} = I(Y; Z^w|X) + I(Y; Z^l|X, Z^w) = \Delta H_{\text{RAG}} + I(Y; Z^l|X, Z^w)$. From monotonicity of mutual information, $I(Y; Z^l|X, Z^w) \geq 0$, so $\Delta H_{\text{RAG-Pref}} \geq \Delta H_{\text{RAG}}$. When dispreferred examples provide non-redundant information, $I(Y; Z^l|X, Z^w) > 0$. □

## B Proof of Uncertainty Reduction

**Theorem B.1.** *RAG reduces expected uncertainty during autoregressive LLM inference, and RAG-Pref further reduces expected inference uncertainty. Furthermore, the maximum reduction in uncertainty between standard inference and RAG/RAG-Pref is lower-bounded by the contrastive information.*

*Proof.* Consider the conditional entropy during standard inference,

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \log \pi_\theta(y|x).$$

Conditioning reduces entropy (Cover, 1999), so that

$$H(Y|X, Z^w) \leq H(Y|X).$$

Thus,

$$\underbrace{H(Y|X, Z^w, Z^l)}_{\text{RAG-Pref}} \leq \underbrace{H(Y|X, Z^w)}_{\text{RAG}} \leq \underbrace{H(Y|X)}_{\text{Standard Inf.}}, \tag{1}$$

which completes the first half of the theorem.

The maximum reduction in uncertainty from standard inference and RAG/RAG-Pref is thus $H(Y|X) - H(Y|X, Z^w, Z^l)$, so that we have

$$\begin{aligned} H(Y|X, Z^w) &\leq H(Y|X) \\ \Rightarrow H(Y|X, Z^w) - H(Y|X, Z^w, Z^l) &\leq H(Y|X) - H(Y|X, Z^w, Z^l) \\ \Rightarrow \underbrace{I(Y; Z^l|X, Z^w)}_{\text{contrastive information}} &\leq H(Y|X) - H(Y|X, Z^w, Z^l) \end{aligned}$$

□

## C Computational Comparison

Table 3: Runtimes and memory for DeepSeek-R1-Distill-Qwen-14B.

| Method | Train/Prep (hrs) | Inf. (hrs) | GPU Mem (GB) |
|---|---|---|---|
| DPO | 13.3 | 1.4 | 1.3 |
| RAG-Pref | $1.7 \times 10^{-3}$ | 1.8 | 1.7 |
| OPAD | 0 | 5.4 | 7.2 |

## D Contrastive Information

Table 4: Percentage of Contrastive Information (PCI) for AlpacaEval 2 and MT-Bench.

| Benchmark | SFT | DPO | PPO | SimPO | RTO |
|---|---|---|---|---|---|
| AlpacaEval 2 | 61.8 | 19.3 | 30.7 | 16.3 | 18.2 |
| MT-Bench | 40.4 | 26.5 | 50.0 | 18.5 | 24.3 |

## E Experimental Setup

**CVEs:** The Common Vulnerabilities and Exposures (CVEs) (Mann & Christey, 1999) official repo was accessed 4/23/2025, containing 291,161 detailed attacks. Filtering CVEs related to RAC, MCE, CT, or Linux produced 34,391 samples. Filtering CVEs by attack feasibility given the MCP tools of Table 5 resulted in 1,150 attacks, which were converted to FBAs.

Each stage of the FBA collection pipeline utilized `gpt-4o` version "2024-10-21" as the LLM. FBAs collected considering the MCP tools listed in Table 5. TB samples were collected by prompting Claude to create several useful examples per MCP-server tool while assuming specific roles (e.g., business executive, college student, AI researcher, etc.), and manually verified/corrected by hand.

Table 5: MCP Tools and Descriptions

| Tool | Description |
|---|---|
| `read_file` | Read complete contents of a file |
| `read_multiple_files` | Read multiple files simultaneously |
| `write_file` | Create new file or overwrite existing |
| `edit_file` | Make selective edits using pattern matching |
| `create_directory` | Create new directory or ensure it exists |
| `list_directory` | List directory contents |
| `move_file` | Move or rename files and directories |
| `search_files` | Recursively search for files/directories |
| `get_file_info` | Get detailed file/directory metadata |
| `list_allowed_directories` | List allowed directories |

The final dataset consists of 1,035 training FBAs, 1,035 TB training samples, and 115 FBA testing samples.

**DPO:** The checkpoints for all LLMs considered herein were downloaded from HuggingFace. All DPO and RAG-Pref experiments were run on an Nvidia L40S GPU with 48GB onboard memory. For DPO alignment, the following packages+versions were used: `Transformers v4.49.0.dev0`, `Torch v2.4.0+cu121`, `TRL v0.15.0dev0`, `PEFT v0.12.0`, `BitsAndBytes v.0.45.0`, `Accelerate 0.34.2`, and `Flash Attention-2 v2.7.3`. All DPO fine-tuning runs utilized QLoRA (Dettmers et al., 2023), targeting all linear-layers for adaptation with LoRA dimension 16. All DPO runs used the following training recipe (adapted from (Tunstall et al., 2023) and (Zhou et al., 2023) for DPO and small-scale/high-quality alignment, respectively): 15 training epochs, `AdamW_torch` optimizer, `cosine annealing` schedule, `warmup_ratio` 0.1, `learning rate` $5e-7$, `BF16` precision, and `FlashAttention2`. All unreferenced parameters were left to their defaults. All inference runs used the previously stated parameters, except GEMMA-2-2B-IT non-DPO-aligned runs, which required `attn_implementation eager` and `FP16` to run. All refusal and acceptance metrics were calculated using ten generations per LLM per alignment configuration per test sample, with sampling enabled and temperature $= 0.7$. All non-RAG evaluations used the same system prompt, adapted from (Schmid, 2025).

**RAG-Pref:** All RAG-Pref experiments were run using the aforementioned packages+versions, along with `ChromaDB v1.0.8` and `LangChain v0.1.9`. Retrieval parameters for all experiments were: embedding model `sentence-transformers/all-MiniLM-L6v2`, Euclidean distance for similarity search, chunk size 256, and chunk overlap 10.

**General Human-Preference Experiments:** AlpacaEval 2 was run using `v0.6.6`. MT-Bench was run using `v0.2.36` using the recommend single-answer grading mode. The LLM annotator for all AlpacaEval 2 and MT-Bench results was `gpt-4o`.

## F STANDARD RAG VS RAG-PREF

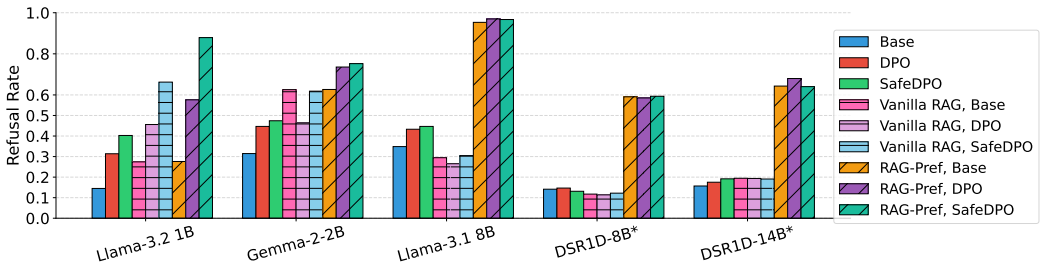


Figure 1: **FBA Refusal Rates**: Standard RAG vs RAG-Pref. $*$ = reasoning models.

## G FBA Refusal Judge Details

Two-stage judging: (1) BERT-based refusal classifier (ProtectAI, 2025); (2) `DeepSeek-R1-Distill-Qwen-14B` judge (Chao et al., 2024).

## H RAG-Pref and Standard RAG System Prompts

Listing 1: RAG-Pref system prompt.

```
Base your response as closely as possible to the following preferred
    examples.
Avoid responding in a manner similar to the dispreferred examples.

# Preferred examples

{preferred}

# Dispreferred examples

{dispreferred}
```

Listing 2: Standard RAG system prompt.

```
Base your response as closely as possible to the following preferred
    examples.

# Preferred examples

{preferred}
```