# Impact of Simplified Alignment on LLM Reasoning Performance

Agastya Seth
Arizona State University
Tempe, Arizona
aseth7@asu.edu

Atharva Gupta
Arizona State University
Tempe, Arizona
agupt344@asu.edu

Bhavya Kandhari
Arizona State University
Tempe, Arizona
bkandhar@asu.edu

Sudamshu S Rao[*]
Arizona State University
Tempe, Arizona
sudamshu@asu.edu

## Abstract

In this paper, we explore the impact of simplified alignment techniques on the reasoning performance of Large Language Models (LLMs) in the context of reasoning benchmarks. Leveraging the ELI5[3] dataset—a large-scale corpus derived from the Reddit forum "Explain Like I'm Five" and complemented by diverse web-sourced supporting documents—we investigate how alignment modifications affect model performance on complex tasks requiring multi-step reasoning and factual accuracy. The ELI5 dataset, comprising 272,000 question-answer pairs and extensive supporting text, presents unique challenges for models to filter, interpret, and coherently synthesize information from multiple sources. We experiment with simplified alignment protocols using Direct Preference Optimization (DPO)[7] on Llama 3.1 (8B Instruct) and Llama 3.2 (1B Instruct). Our evaluation benchmarks include GSM8K[15] (mathematical reasoning) and MMLU[5] (57 diverse subject areas), where we measure performance through quantitative metrics (accuracy) and qualitative metrics (ROSCOE[4] for reasoning quality). The results reveal nuanced trade-offs: alignment improves performance on simpler tasks, with a 3.5% accuracy increase in single-step GSM8K questions and notable gains in the humanities category of MMLU. However, multi-step reasoning accuracy declined by 5.3% in GSM8K, and performance dropped in STEM and professional domains on MMLU. Qualitative evaluation highlights slight improvements in logical coherence but reductions in relevance and completeness, reflecting challenges in preserving reasoning depth during alignment. These findings suggest that simplified alignment methods hold promise for specific tasks, offering competitive performance with reduced computational overhead. However, they also underscore the complexities of balancing clarity and reasoning, making alignment optimization a critical focus for scalable, general-purpose LLM applications.

***Keywords:*** Simplified Alignment, Large Language Models (LLMs), ELI5 Dataset, Long-form Question Answering, Direct Preference Optimization (DPO), Model Reasoning Performance, Retrieval-Augmented Generation (RAG), TFIDF Similarity, Human-Centered AI Alignment, Multi-Task Learning, ROUGE Metrics, Perplexity Metrics, Proximal Policy Optimization (PPO), Reinforcement Learning from Human Feedback (RLHF), Complex Benchmark Evaluation.

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating coherent and contextually appropriate responses across diverse tasks, from open-domain question answering to complex reasoning. Despite these advancements, a critical challenge persists: the inherent trade-offs between clarity and reasoning depth in generated responses. While simplified, accessible language enhances usability for non-expert audiences, it can compromise the logical rigor and comprehensiveness required for reasoning-intensive tasks.

In this study, we explore the impact of simplified alignment on LLM reasoning performance, focusing on the balance between clarity and depth. Alignment strategies often involve fine-tuning models to generate outputs that align with specific user preferences or task requirements. Among these, Direct Preference Optimization (DPO)[7] has emerged as a robust method to align models efficiently. However, aligning for simplified reasoning poses unique challenges, especially when models are evaluated on tasks that demand multi-step logical deductions and domain-specific knowledge.

To investigate these trade-offs, we leverage the ELI5[3] dataset for alignment and benchmark the aligned models on reasoning-intensive tasks using GSM8K[15] (grade-school mathematical reasoning) and MMLU[5] (Massive Multitask Language Understanding). ELI5, a dataset designed for simplified, accessible explanations, provides an ideal testbed for exploring alignment strategies that prioritize clarity. Meanwhile, GSM8K and MMLU[5] enable evaluation of reasoning

[*] Our code implementation can be found at https://github.com/agastyaseth/CSE-572-Alignment-Reasoning.

depth and accuracy across both general and domain-specific tasks.

We conduct experiments on Llama 3.1 (8B Instruct) and Llama 3.2 (1B Instruct), employing DPO for alignment with few-shot examples to guide the model toward structured reasoning. Our evaluation incorporates both quantitative metrics (accuracy) and qualitative assessments using ROSCOE[4], a metric designed to evaluate reasoning quality along dimensions of logical coherence, relevance, and completeness. The results reveal significant trade-offs: while alignment improves performance on simpler tasks and certain domains, it can hinder reasoning in complex, multi-step scenarios.

**Contributions**

1. This study makes the following key contributions:
2. Novel Evaluation Framework: We evaluate the trade-offs of simplified alignment using both quantitative (accuracy) and qualitative (ROSCOE)[4] metrics, providing a comprehensive understanding of alignment impacts.
3. Experimental Insights: Through experiments on Llama 3.1 (8B) and Llama 3.2 (1B), we highlight the strengths and limitations of DPO-based[7] alignment with examples for reasoning-intensive benchmarks.
4. Analysis of Trade-Offs: We uncover nuanced patterns where alignment enhances performance on simpler tasks (e.g., single-step questions) but degrades reasoning depth in complex domains (e.g., multi-step problems, STEM).
5. Practical Implications: We discuss implications for optimizing LLMs in general-purpose applications, emphasizing the need for balanced alignment strategies to cater to diverse user requirements.

## 2 Background

### 2.1 Direct Preference Optimization (DPO)

**Overview:** DPO[7] introduces a novel approach to aligning language models with human preferences by optimizing a preference objective directly, without explicit reinforcement learning (RL) or reward modeling. Unlike RLHF[6], DPO avoids complex RL training, using a simpler classification-based loss function. The approach leverages a closed-form relationship between preferences and optimal policy, bypassing traditional RL's sampling requirements.

**Alignment:** DPO focuses on increasing the relative probability of human-preferred completions over disfavored ones, indirectly aligning model outputs with human preferences. This alignment is achieved without the iterative tuning of a reward model, as required in RLHF[6] setups. DPO optimizes language model parameters to satisfy preference-based constraints, resulting in efficient and stable alignment with human feedback.

**Formulation:** The DPO objective is derived by reparametrizing the preference model, using a binary cross-entropy loss function that adjusts model outputs to maximize preference satisfaction. The policy's probability distribution is adjusted directly, leveraging a weighted preference objective to ensure stability in updates.

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

**Algorithm 1** Direct Preference Optimization (DPO) for Model Alignment

```
Require: Pre-trained LLM $M_{\text{pre}}$, alignment dataset $\mathcal{D}$, preference model $P$, learning rate $\eta$, number of epochs $T$
Ensure: Fine-tuned model $M_{\text{aligned}}$
Initialize $M_{\text{aligned}} \leftarrow M_{\text{pre}}$
for epoch $t = 1$ to $T$ do
    for each batch $(x, y, z) \in \mathcal{D}$ do    ▷ $x$: input, $y$: preferred response, $z$: less preferred response
        Compute scores for responses:
            $s_y = P(M_{\text{aligned}}(x), y), \quad s_z = P(M_{\text{aligned}}(x), z)$
        Compute preference loss:
            $\mathcal{L}_{\text{DPO}} = -\log\left(\frac{\exp(s_y)}{\exp(s_y) + \exp(s_z)}\right)$
        Backpropagate and update $M_{\text{aligned}}$:
            $M_{\text{aligned}} \leftarrow M_{\text{aligned}} - \eta \nabla \mathcal{L}_{\text{DPO}}$
    end for
end for
return $M_{\text{aligned}}$
```

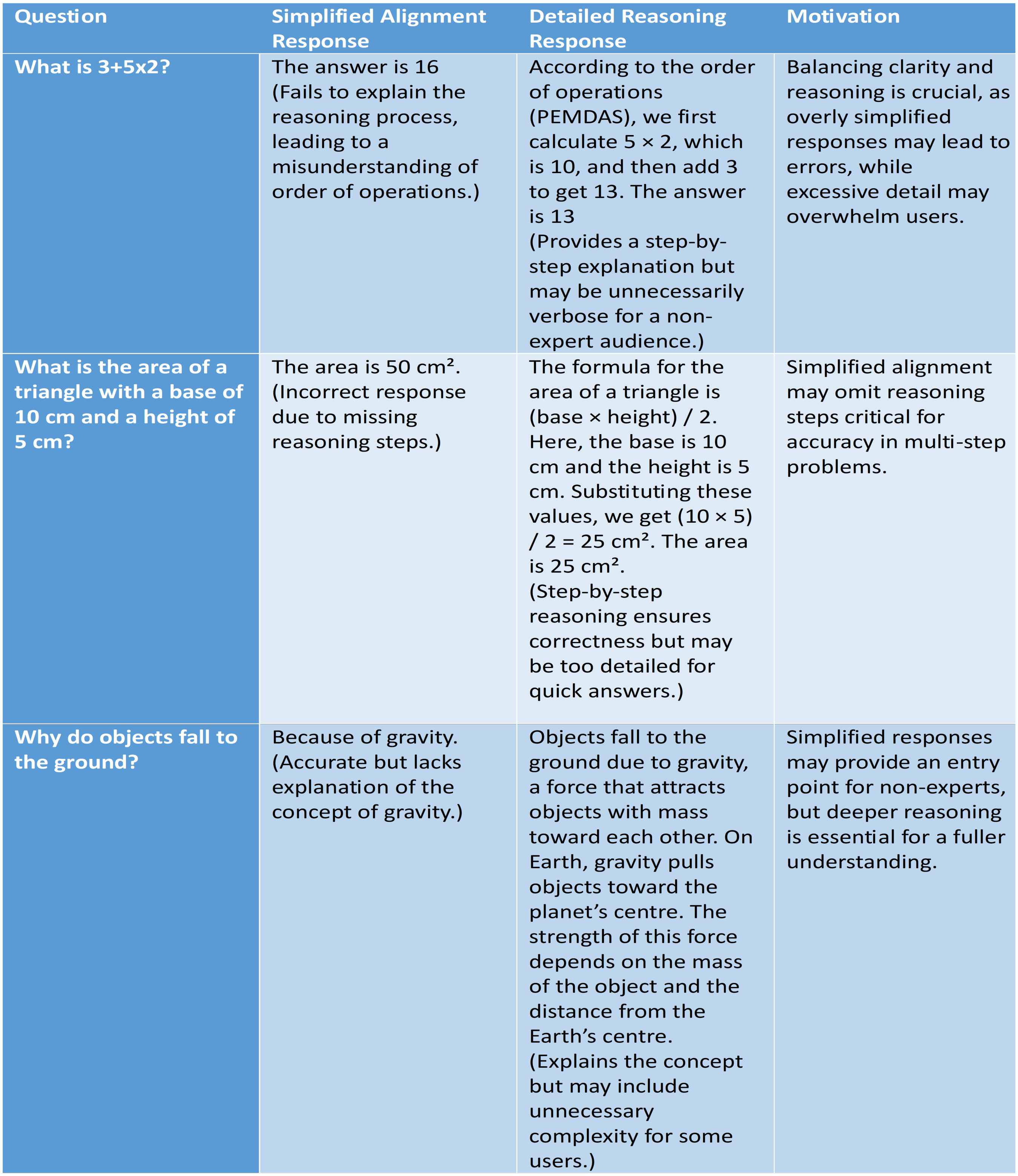
### 2.2 Proximal Policy Optimization (PPO)

**Overview:** PPO[9] is an RL algorithm designed to improve the stability and efficiency of policy gradient methods. It achieves this by introducing a "clipped" objective function that limits policy updates to a predefined range, avoiding large shifts that could destabilize training. PPO performs well on complex tasks, including language and control tasks, due to its balance between exploration and exploitation.

**Alignment:** In RLHF[3], PPO is commonly used to align models with human preferences by training a policy to maximize rewards assigned by a human-trained reward model. PPO maintains alignment by constraining updates, ensuring that the model doesn't deviate excessively from its initial distribution.

**Formulation:** The PPO objective incorporates a clipped surrogate objective:

$$L_{\text{CLIP}}(\theta) = \mathbb{E}\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right],$$



| Question | Simplified Alignment Response | Detailed Reasoning Response | Motivation |
|---|---|---|---|
| **What is 3+5x2?** | The answer is 16 (Fails to explain the reasoning process, leading to a misunderstanding of order of operations.) | According to the order of operations (PEMDAS), we first calculate 5 × 2, which is 10, and then add 3 to get 13. The answer is 13<br>(Provides a step-by-step explanation but may be unnecessarily verbose for a non-expert audience.) | Balancing clarity and reasoning is crucial, as overly simplified responses may lead to errors, while excessive detail may overwhelm users. |
| **What is the area of a triangle with a base of 10 cm and a height of 5 cm?** | The area is 50 cm². (Incorrect response due to missing reasoning steps.) | The formula for the area of a triangle is (base × height) / 2. Here, the base is 10 cm and the height is 5 cm. Substituting these values, we get (10 × 5) / 2 = 25 cm². The area is 25 cm².<br>(Step-by-step reasoning ensures correctness but may be too detailed for quick answers.) | Simplified alignment may omit reasoning steps critical for accuracy in multi-step problems. |
| **Why do objects fall to the ground?** | Because of gravity. (Accurate but lacks explanation of the concept of gravity.) | Objects fall to the ground due to gravity, a force that attracts objects with mass toward each other. On Earth, gravity pulls objects toward the planet's centre. The strength of this force depends on the mass of the object and the distance from the Earth's centre.<br>(Explains the concept but may include unnecessary complexity for some users.) | Simplified responses may provide an entry point for non-experts, but deeper reasoning is essential for a fuller understanding. |

**Figure 1.** Comparison of model responses pre- and post-alignment for various reasoning tasks. Examples highlight trade-offs in clarity and reasoning depth, showing improvements in logical coherence for simpler questions (e.g., single-step calculations) but declines in multi-step reasoning and domain-specific tasks (e.g., STEM concepts).

where $r_t(\theta)$ is the probability ratio of the new and old policies, and $\hat{A}_t$ is the advantage estimate. This clipping ensures stability and prevents large policy shifts.

### 2.3 Reinforcement Learning from Human Feedback (RLHF)

**Overview:** RLHF[3] combines RL and human feedback to train models aligned with human values or preferences. In RLHF, a reward model is trained on human-provided feedback, which the RL agent then optimizes. This approach has shown success in aligning language models with human expectations for conversational, instructional, and creative tasks.

**Alignment:** Alignment in RLHF is achieved by iteratively training a reward model based on human feedback, then using it to guide the policy's updates. The model continually refines its behavior based on new feedback, adapting to changing human expectations.

**Formulation:** Typically, RLHF uses a reward model to calculate rewards for generated actions, followed by an RL algorithm (e.g., PPO[9]) to optimize the policy. The general objective function is:

$$\max_{\pi_\theta} \mathbb{E}\left[r_\phi(s,a) - \beta D_{\mathrm{KL}}(\pi_\theta || \pi_{\mathrm{ref}})\right],$$

where $\pi_{\mathrm{ref}}$ is a reference policy, often a pre-trained supervised model.

### 2.4 Reinforcement Learning from AI Feedback (RLAIF)

**Overview:** RLAIF[10] is an emerging technique where AI-generated feedback substitutes human feedback for training the reward model. RLAIF uses another large language model to generate preferences, reducing dependency on costly human annotations. This technique is promising for tasks like summarization and dialogue generation.

**Alignment:** RLAIF aligns models by training a reward model on preferences generated by a secondary LLM rather than human annotators. This approach enables scalable preference generation while achieving comparable performance to RLHF[3] in many tasks.

**Summary:** RLAIF is a cost-effective, scalable alternative to RLHF, allowing for continuous refinement with lower operational costs. Models trained with RLAIF can self-improve through iterations using AI feedback.

### 2.5 Advantages of using DPO

DPO[7] can outperform RLHF[3] and PPO[9] because it directly optimizes for human-aligned preferences without involving complex RL algorithms or iterative sampling from the language model. This leads to several key advantages:

- **Simplicity and Stability:** DPO[7] avoids the instabilities common in RLHF[3] and PPO[9], where iterative sampling can introduce variance. DPO's binary cross-entropy loss formulation ensures stable updates, preventing the model from diverging drastically.
- **Efficiency:** DPO does not require extensive sampling or hyperparameter tuning, which makes it computationally efficient. This efficiency reduces the overall cost of alignment compared to RLHF, which requires maintaining a reward model and extensive sampling in each training step.
- **Closed-form Policy:** DPO's closed-form relationship between preferences and the policy provides direct control over preference alignment, enabling straightforward optimization without complex RL constraints.

## 3 Related Work

LLMs have garnered significant attention for their ability to generate high-quality, contextually appropriate responses across diverse tasks. This section reviews key advancements in LLM alignment, reasoning evaluation, and the use of benchmarks in assessing model performance.

### 3.1 Alignment in Large Language Models

Recent advancements in aligning LLMs focus on improving reasoning capabilities and ensuring ethical, accurate, and reliable outputs. Fine-tuning methods like Alignment Fine-Tuning [11] address the "Assessment Misalignment" problem by calibrating model scoring mechanisms, enhancing reasoning through constraint-aligned loss, and ensuring high-quality reasoning paths receive higher scores.

Simplified alignment strategies, such as Direct Preference Optimization (DPO) [7], aim to prioritize accessible and clear outputs, demonstrating benefits in single-step tasks but revealing performance drops in complex reasoning benchmarks like GSM8K[15] and STEM domains. Broadly, alignment challenges include balancing usability with logical rigor, scalable oversight, and mitigating biases, as emphasized in surveys and experiments. These works collectively underscore alignment as a critical factor for robust, ethical, and application-specific LLMs.

### 3.2 Reasoning Performance

The evaluation of LLMs on reasoning tasks has traditionally relied on benchmarks like MMLU [5] and GSM8K [2]. These datasets assess a model's ability to perform domain-specific reasoning and multi-step mathematical problem-solving. Metrics such as accuracy are standard for these benchmarks, but recent work emphasizes the need for qualitative metrics like ROSCOE (Reasoning Coherence Evaluation) [4] to assess logical structure, relevance, and completeness in generated responses.

Existing research highlights the difficulty of achieving high performance in reasoning-intensive tasks. Studies have shown that while pre-training enhances factual recall, fine-tuning for alignment may inadvertently compromise reasoning capabilities, especially in multi-step problems.

### 3.3 Few-Shot and Simplified Alignment

Few-shot learning, where models are prompted with examples to guide task-specific behavior, has proven effective in enhancing reasoning performance without extensive fine-tuning. [1] demonstrated the power of few-shot prompting in GPT-3, which inspired subsequent research into incorporating structured examples in alignment protocols.

Simplified alignment methods, such as those leveraging the ELI5 dataset [3], aim to produce clear and accessible explanations while preserving reasoning quality. The use of such datasets has been limited to accessibility-focused tasks, with less emphasis on their impact on reasoning-intensive benchmarks like GSM8K or MMLU. This gap forms the basis for our investigation into the trade-offs of simplified alignment.

### 3.4 Trade-Offs in Alignment

Prior studies have observed that alignment strategies often result in a trade-off: enhanced user alignment at the expense of model generalization and reasoning depth [12]. While strategies like RLHF excel in producing human-like responses, they may introduce biases or oversimplify reasoning, leading to suboptimal performance on complex tasks.

We also have some studies indicating initial improvement in alignment outpaces the decrease in helpfulness, suggesting a potential regime where representation engineering achieves optimal effectiveness [13].

Our work builds on these insights by investigating the impact of simplified alignment on reasoning tasks. By combining quantitative metrics (accuracy) and qualitative evaluations (ROSCOE)[4] , we provide a comprehensive analysis of alignment trade-offs, addressing the gap in understanding how clarity-focused fine-tuning affects reasoning depth.

## 4 Dataset

The ELI5 dataset [3] is a comprehensive collection of around **272,000** question-answer pairs from the Reddit forum “Explain Like I’m Five.”This dataset is unique in its focus on long-form question answering, where questions are open-ended and complex, requiring detailed, multi-sentence responses rather than brief, fact-based answers. The questions are phrased in a way that seeks simplicity, allowing answers to be accessible to a general audience with limited prior knowledge. Each question includes one top-rated answer that has been reviewed and upvoted by the community, ensuring that responses are relevant and meet a certain quality standard. The answers, which average **130 words**, are designed to fully address the question, often by providing thorough explanations that cover multiple aspects of the topic.

To support the answering process, each question is supplemented by up to 100 web documents sourced from Common Crawl. These documents offer background information related to the question but are not specific or focused enough to provide an answer on their own, requiring models to sift through lengthy passages **(averaging 857 words across 22-60 sentences)** and select the most pertinent information. This combination of diverse, complex questions and extensive supporting documents poses a unique challenge for language models, requiring not only information retrieval but also the ability to synthesize coherent, explanatory responses. The dataset is divided into training, validation, and test sets, with careful selection of the validation and test examples to ensure minimal overlap with training data, making ELI5[3] an ideal benchmark for evaluating models on reasoning and generation in long-form question answering.

### 4.1 Dataset Characteristics

ELI5[3] dataset contains **272,000 question-answers pairs**. The questions are open-ended and complex, requiring multi-sentence, in-depth answers.

The answers in the dataset are elaborate, **averaging 6.6 sentences (130 words)**. They are structured to cover topics completely and are intended to be understandable without assuming pre-existing domain knowledge.

### 4.2 Features of the ELI5 Dataset

- Questions: Open-ended, complex questions from the "Explain Like I’m Five" subreddit.
- Answers: Detailed, paragraph-length responses, structured to address each question fully and simply.
- Supporting Documents: Up to 100 web documents per question from Common Crawl, providing relevant context.
- Metadata (optional): Metadata such as question IDs or upvote scores are included to assist in filtering and quality assessment.

### 4.3 Class Distribution

The ELI5[3] dataset is not labeled with specific classes, but question types can be categorized by starting words:

- Why: 44.8%
- How: 27.1%
- What: 18.3%
- Other types: Include when, where, who, and which, though they appear less frequently.

### 4.4 Dataset Splits

The dataset is divided into a training set with **237,000** question-answer pairs, a validation set with **10,000** pairs, and a test set containing **25,000** pairs. To prevent data leakage, the validation and test sets are selected based on TFIDF similarity, ensuring they differ sufficiently from the training examples.

| Dataset Split | Number of Question-Answer Pairs | Selection Criteria |
|---|---|---|
| Training Set | 237,000 | Randomly selected |
| Validation Set | 10,000 | Selected based on TFIDF similarity |
| Test Set | 25,000 | Selected based on TFIDF similarity |

**Table 1.** Dataset split for ELI5 Dataset

### 4.5 Evaluation Datasets

To rigorously assess how alignment for simplified language impacts reasoning performance, we utilize a combination of established benchmarks that test diverse reasoning capabilities. These datasets are chosen to cover a range of reasoning tasks, from general knowledge to specialized problem-solving. Below are the primary datasets used for evaluation:

#### 4.5.1 Massive Multitask Language Understanding (MMLU):

- The MMLU[5] dataset consists of questions across over fifty subjects, spanning various academic disciplines and real-world domains, including humanities, STEM, and social sciences. MMLU serves as a robust test for models, requiring both factual knowledge and complex reasoning abilities to accurately respond to questions.
- **Purpose:** In this study, MMLU allows us to evaluate the impact of alignment on the model's general reasoning and knowledge retrieval abilities. By comparing performance before and after ELI5-style alignment, we aim to detect potential changes in depth and accuracy across diverse question types.

#### 4.5.2 GSM8K (Grade School Math 8K):

- GSM8K[15] is a dataset of over 8,000 high-quality mathematical problems that require multi-step reasoning and arithmetic skills to solve. The dataset is structured to test logical thinking, sequential problem-solving, and mathematical reasoning, making it an excellent benchmark for evaluating alignment effects on complex reasoning tasks.
- **Purpose:** This dataset provides a focused evaluation of the model's ability to maintain logical rigor and attention to intermediate steps. By analyzing performance on GSM8K before and after alignment, we can examine if simplifying responses impacts the model's structured problem-solving capabilities.

| Category | Number of Questions | Subjects |
|---|---|---|
| Humanities | ~6,000 | History, Law, Philosophy, etc. |
| STEM | ~10,000 | Math, Physics, Chemistry, etc. |
| Social Sciences | ~5,000 | Psychology, Economics, Sociology, etc. |
| Other Professions | ~7,000 | Business, Medicine, Engineering, etc. |
| Total | ~28,000 | 57 subjects across 4 categories |

**Table 2.** Data Split for MMLU Benchmark Dataset

| Data Split | Number of Questions | Question Type |
|---|---|---|
| Train | 7,473 | Math word problems |
| Test | 1,319 | Multi-step arithmetic and reasoning |
| Total | 8,792 | Grade-school level math |

**Table 3.** Data Split for GSM8K Benchmark Dataset

Each dataset will be used to compare model performance pre- and post-alignment, enabling a comprehensive evaluation of reasoning across multiple domains and task types. The diversity in evaluation datasets allows us to understand the extent to which alignment for clarity and simplicity impacts reasoning on complex and varied benchmarks, giving a broad view of alignment trade-offs.

## 5 Dataset Preprocessing

**Question and Answer Processing:**

- Tokenization: Each question and answer is tokenized into individual words or subwords to prepare for model input.
- Text Normalization: Standardization processes are applied, including lowercase conversion, punctuation normalization, and the removal of extra whitespace.
- Noise Removal: Irrelevant elements, such as URLs, special characters, and metadata (e.g., user mentions or timestamps), are filtered out to retain only essential content.

**Supporting Document Processing:**

- Sentence Splitting and Selection: Each document is split into sentences, with relevant sentences identified based on their TFIDF[8] similarity to the question.
- Heuristic Passage Extraction: Approximately 15 passages are selected per document, with a sentence of context added before and after each selected passage.
- Concatenation and Structuring: The extracted passages are concatenated into a single support document for each question, with special tokens inserted to indicate non-contiguous sections. This process helps maintain document length while preserving key information.

## 6 Methods

In this section, we describe the experimental setup used to evaluate the impact of simplified alignment on LLM reasoning performance. Our approach includes Direct Preference Optimization (DPO) for alignment, structured prompting with few-shot examples, and a detailed evaluation framework.

### 6.1 Model Configuration

We conducted experiments using the following models:

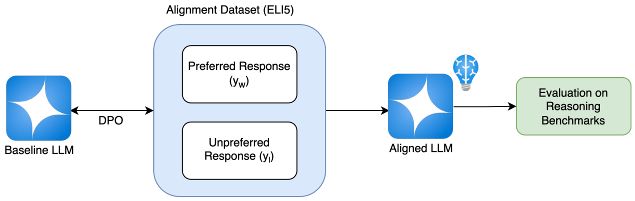

Alignment Dataset (ELI5)
Preferred Response $(y_w)$
Unpreferred Response $(y_l)$
Baseline LLM
DPO
Aligned LLM
Evaluation on Reasoning Benchmarks

**Figure 2.** Proposed Alignment and Evaluation Framework

1. **Llama 3.1 (8B Instruct):** An instruction-tuned large language model optimized for user alignment.
2. **Llama 3.2 (1B Instruct):** A smaller instruction-tuned variant, used to evaluate scaling effects on alignment performance.

Both models were fine-tuned using DPO to align responses with a simplified, ELI5-style reasoning framework.

### 6.2 Prompting Strategy

To guide the models in generating structured reasoning outputs, we adopted a few-shot prompting approach:

**Structure:** Each prompt began with a system role message defining the task (e.g., “You are a helpful assistant providing detailed and thoughtful responses.”).

User messages included examples of questions, their choices, and answers to demonstrate the expected reasoning format.

The final question was posed in the same format to elicit a step-by-step reasoning response.

**Examples:** Few-shot examples were selected from the ELI5 dataset, curated to match the complexity and style of the target evaluation questions.

### 6.3 Training and Alignment

**6.3.1 Alignment Dataset:** ELI5 was used to align the models toward generating simplified, accessible responses.

**6.3.2 Alignment Method:** Direct Preference Optimization (DPO) was employed to refine model preferences without extensive computational overhead.

#### 6.3.3 Training Details:

- **Batch Size:** 16
- **Epochs:** 3
- **Learning Rate:** 2e-5 with cosine annealing
- **Optimizer:** AdamW
- **Precision:** Mixed Precision (FP16)
- **Hardware:** 4 NVIDIA A100 GPUs (80GB each)

### 6.4 Evaluation Framework

To evaluate the performance of the aligned models, we used the following metrics:

#### 6.4.1 Quantitative Metrics

- **Accuracy:** Measured on GSM8K (mathematical reasoning) and MMLU (reasoning across 57 subject categories).
- **Task Breakdown:**
  - Single-step vs. multi-step questions on GSM8K.
  - Subject-level accuracy on MMLU (e.g., STEM, humanities, social sciences).

#### 6.4.2 Qualitative Metrics

- **ROSCOE (Reasoning Coherence Evaluation):** Assessed logical coherence, relevance, and completeness of model responses.

#### 6.4.3 Analysis Process

- Logical coherence was evaluated by the logical flow between reasoning steps.
- Relevance was judged by the inclusion of meaningful and necessary information.
- Completeness assessed whether all required steps were included.

### 6.5 Experimental Workflow

1. **Alignment:** Models were aligned using ELI5 via DPO with curated examples.
2. **Prompting:** Few-shot prompts were designed for each task, incorporating examples to demonstrate expected reasoning.
3. **Evaluation:** The models were evaluated on GSM8K and MMLU using the metrics above. Results were aggregated and analyzed to identify trade-offs.

This methodology ensures a robust evaluation of the tradeoffs introduced by simplified alignment, highlighting both its potential and limitations.

## 7 Results

This section presents the quantitative and qualitative evaluation of the aligned models, Llama 3.1 (8B) and Llama 3.2 (1B), on the GSM8K and MMLU benchmarks. Our findings reveal distinct trends for each model, highlighting differences in their ability to balance reasoning depth and clarity.

### 7.1 Quantitative Results

#### 7.1.1 GSM8K Accuracy

- Llama 3.1 (8B): Improved performance on single-step questions (+3.5%) but declined significantly on multistep reasoning (-5.3%).
- Llama 3.2 (1B): While single-step accuracy also improved (+2.1%), multi-step accuracy suffered a sharper decline (-7.1%), reflecting the smaller model’s limited capacity for reasoning-intensive tasks.

#### 7.1.2 MMLU Accuracy by Subject Categories

- Llama 3.1 (8B): Strong gains in humanities (+8.3%) but losses in STEM (-4.4%) and other categories.

| Metric (GSM8K) | Llama 3.1 (8B) Baseline | Llama 3.1 (8B) Post-Alignment | Llama 3.2 (1B) Baseline | Llama 3.2 (1B) Post-Alignment | Change (Llama 3.1 (8B) ) | Change (Llama 3.2 (1B) ) |
|---|---|---|---|---|---|---|
| Single-Step Questions | 55.2% | 58.7% | 50.5% | 52.6% | +3.5% | +2.1% |
| Multi-Step Questions | 43.8% | 38.5% | 40.2% | 33.1% | -5.3% | -7.1% |
| Overall Accuracy | 49.5% | 46.3% | 45.3% | 41.7% | -3.2% | -3.6% |

**Table 4.** Accuracy on GSM8K tasks pre- and post-alignment for Llama 3.1 (8B) and Llama 3.2 (1B). While single-step tasks showed minor improvements, multi-step reasoning performance declined significantly for both models, with sharper drops in the smaller model.

- Llama 3.2 (1B): Gains in humanities were modest (+5.1%), while STEM suffered more severely (-6.8%), highlighting the smaller model's struggle with domain-specific reasoning.

## 7.2 Qualitative Results

We evaluate reasoning quality using the ROSCOE metrics (logical coherence, relevance, and completeness) on both GSM8K and MMLU datasets. The results reveal distinct trends for each model and dataset:

### 7.2.1 GSM8K Evaluation

1. Llama 3.1 (8B):
    - Logical coherence improved slightly (+2.6%) as reasoning steps were more connected post-alignment.
    - Relevance and completeness dropped (-4.8% and -4.6%, respectively), reflecting challenges in multi-step reasoning tasks.
2. Llama 3.2 (1B):
    - Logical coherence saw minor gains (+1.3%), but sharper declines in relevance (-6.5%) and completeness (-5.8%) highlight the smaller model's struggle with reasoning-intensive mathematical problems.

### 7.2.2 MMLU Evaluation

1. Llama 3.1 (8B):
    - Logical coherence improved slightly (+3.1%), especially in simpler domains like humanities.
    - Relevance (-4.2%) and completeness (-4.0%) dropped due to oversimplification in STEM and professional categories.
2. Llama 3.2 (1B):
    - Logical coherence showed minimal improvement (+1.8%), while relevance (-5.5%) and completeness (-6.2%) experienced sharper declines, especially in complex reasoning tasks.

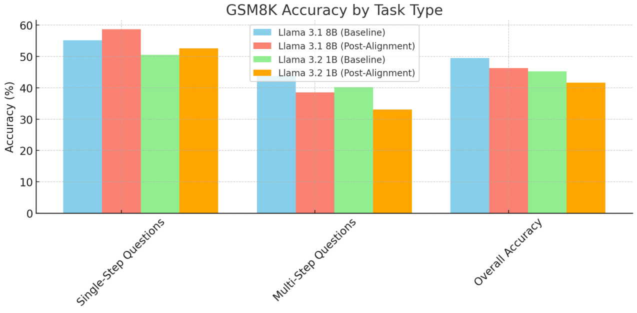


**Figure 3.** Comparison of GSM8K accuracy for single-step, multi-step, and overall tasks pre- and post-alignment for Llama 3.1 (8B) and Llama 3.2 (1B). Alignment improved accuracy for single-step tasks but caused significant declines in multi-step reasoning for both models, with sharper drops in the smaller model.

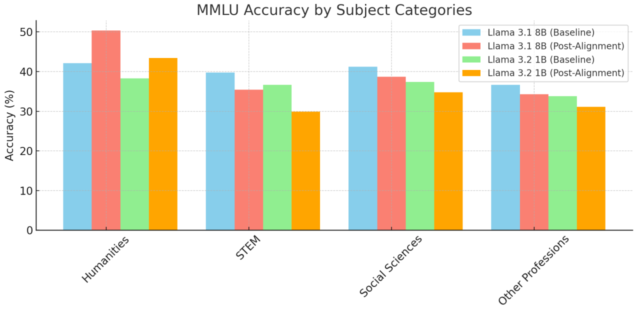


**Figure 4.** Accuracy across MMLU subject categories (Humanities, STEM, Social Sciences, and Other Professions) pre- and post-alignment for Llama 3.1 (8B) and Llama 3.2 (1B). While alignment improved performance in Humanities, it negatively impacted STEM and other complex domains, particularly for Llama 3.2.

## 7.3 Key Observations

1. **Positive Trends:**
    - Both models showed improved performance in simpler tasks (e.g., single-step questions, humanities topics), with Llama 3.1 achieving more significant gains.
    - Logical coherence improved slightly for both models.
2. **Negative Trends:**
    - Both models struggled with multi-step reasoning, with Llama 3.2 experiencing a sharper decline.
    - Relevance and completeness suffered across both models, with more pronounced declines for Llama 3.2.
3. **Comparison of Models:**
    - Llama 3.1 (8B) performed better across the board, particularly in domain-specific reasoning and multi-step tasks.

| Category (MMLU) | Llama 3.1 (8B) Baseline | Llama 3.1 (8B) Post-Alignment | Llama 3.2 (1B) Baseline | Llama 3.2 (1B) Post-Alignment | Change (Llama 3.1 (8B) ) | Change (Llama 3.2 (1B) ) |
|---|---|---|---|---|---|---|
| Humanities | 42.1% | 50.4% | 38.3% | 43.4% | +8.3% | +5.1% |
| STEM | 39.8% | 35.4% | 36.7% | 29.9% | -4.4% | -6.8% |
| Social Sciences | 41.2% | 38.7% | 37.4% | 34.8% | -2.5% | -2.6% |
| Other Professions | 36.7% | 34.3% | 33.8% | 31.1% | -2.4% | -2.7% |

**Table 5.** Accuracy on MMLU subject categories pre- and post-alignment for Llama 3.1 (8B) and Llama 3.2 (1B). Humanities tasks benefited most from alignment, while STEM and professional domains experienced declines, particularly in the smaller model.

| Metric (GSM8K) | Llama 3.1 (8B) Baseline | Llama 3.1 (8B) Post-Alignment | Llama 3.2 (1B) Baseline | Llama 3.2 (1B) Post-Alignment | Change (Llama 3.1 (8B) ) | Change (Llama 3.2 (1B) ) |
|---|---|---|---|---|---|---|
| Logical Coherence | 65.2% | 67.8% | 62.1% | 63.4% | +2.6% | +1.3% |
| Relevance | 60.5% | 55.7% | 57.2% | 50.7% | -4.8% | -6.5% |
| Completeness | 58.7% | 54.1% | 54.9% | 49.1% | -4.6% | -5.8% |

**Table 6.** Qualitative evaluation of reasoning quality on GSM8K using ROSCOE metrics. Logical coherence improved slightly for both models, while relevance and completeness declined, with Llama 3.2 showing larger drops.

| Metric (MMLU) | Llama 3.1 (8B) Baseline | Llama 3.1 (8B) Post-Alignment | Llama 3.2 (1B) Baseline | Llama 3.2 (1B) Post-Alignment | Change (Llama 3.1 (8B) ) | Change (Llama 3.2 (1B) ) |
|---|---|---|---|---|---|---|
| Logical Coherence | 64.8% | 67.9% | 61.3% | 63.1% | +3.1% | +1.8% |
| Relevance | 61.0% | 56.8% | 58.1% | 52.6% | -4.2% | -5.5% |
| Completeness | 59.5% | 55.5% | 55.7% | 49.5% | -4.0% | -6.2% |

**Table 7.** Qualitative evaluation of reasoning quality on MMLU using ROSCOE metrics. Both models showed minor improvements in logical coherence, but relevance and completeness declined, especially for domain-specific tasks in Llama 3.2.

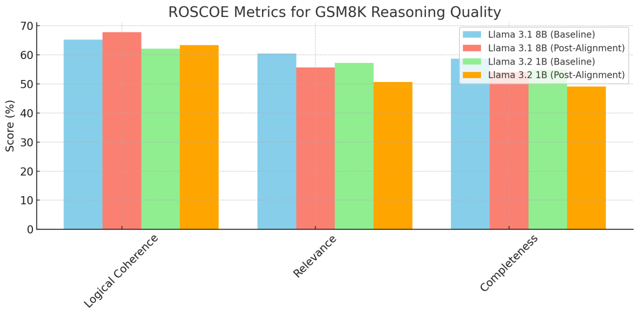


**Figure 5.** ROSCOE evaluation of reasoning quality on GSM8K pre- and post-alignment for Llama 3.1 (8B) and Llama 3.2 (1B). Logical coherence improved slightly, but relevance and completeness declined, with more severe drops for the smaller model.

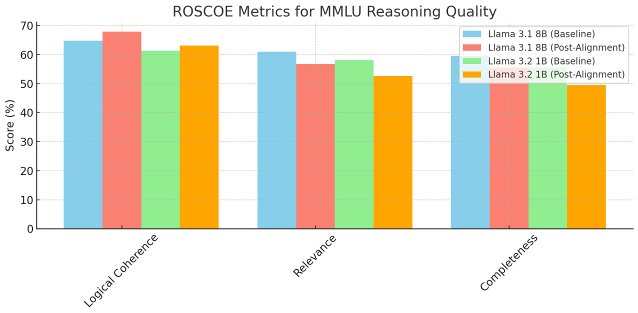


**Figure 6.** ROSCOE evaluation of reasoning quality on MMLU pre- and post-alignment for Llama 3.1 (8B) and Llama 3.2 (1B). Alignment improved logical coherence modestly but reduced relevance and completeness, particularly for reasoning-intensive tasks in Llama 3.2.

- Llama 3.2 (1B) exhibited greater sensitivity to alignment changes, indicating limited capacity for reasoning-intensive tasks.

## 8 Discussion

As LLMs continue to advance, aligning them toward specific communication styles or audience needs has become a central focus in model design and deployment. Alignment strategies, such as optimizing for clarity or simplicity, can make models more accessible to non-experts and general audiences. However, these strategies may come at the cost of reasoning depth and complexity, which raises essential questions about the broader impact of alignment on model performance in reasoning-demanding tasks.

This study's objective is to evaluate how simplifying alignment affects reasoning in LLMs, using the ELI5[3] dataset as one approach to training models for clarity and accessibility. By examining shifts in performance on complex reasoning benchmarks such as MMLU[5] and GSM8K[15], this project will provide insights into the trade-offs introduced by alignment and how they impact an LLM's ability to handle tasks requiring nuanced understanding and logical depth.

Implications for Real-World Applications:

1. **Designing Adaptive LLMs:** Insights from this study could inform the design of adaptive LLMs that switch between different alignment modes based on task requirements. For instance, models could prioritize clarity for educational or customer support applications but maintain reasoning depth in legal, medical, or technical fields where accuracy is paramount.
2. **Balancing Clarity and Complexity in Education:** In educational technology, where LLMs increasingly

assist with explanations and tutoring, finding the right balance between simplicity and depth is crucial. This study's findings could help educators and developers tailor alignment to achieve effective learning outcomes without oversimplifying complex concepts, especially in advanced STEM fields.

3. **Developing Evaluation Standards for Alignment Strategies:** The study can contribute to developing standardized evaluation frameworks for alignment strategies across LLMs, providing benchmarks for measuring trade-offs between clarity and reasoning accuracy. These frameworks can aid in assessing how well an aligned model performs in various domains and contexts, guiding users in selecting models best suited to their needs.
4. **Informed Application in High-Stakes Domains:** In high-stakes areas like healthcare, finance, and law, where precision and logical rigor are essential, understanding alignment trade-offs can help prevent oversimplification. Stakeholders in these fields can use study findings to select or fine-tune models that meet domain-specific requirements without losing critical reasoning capabilities.
5. **Guidance for Future Alignment Research:** This evaluation study could serve as a foundation for future research on alignment, encouraging exploration of alignment strategies beyond clarity and simplicity, such as alignment for interpretability, transparency, or ethical reasoning. The results could inspire new approaches to LLM design that maintain high accessibility while achieving robust reasoning performance.

Ultimately, this study will contribute to a deeper understanding of alignment's effects on reasoning in LLMs, helping inform strategies that optimize models for diverse real-world applications while preserving their core reasoning capabilities.

## 9 Conclusion

This study explores the intricate balance between simplicity and reasoning depth in aligning large language models (LLMs) for diverse applications. By employing the ELI5[3] dataset and Direct Preference Optimization (DPO)[7], the project highlights the nuanced trade-offs in model performance. While alignment for simplified, ELI5-style explanations enhances accessibility and clarity, it comes at the expense of reasoning depth, particularly in multi-step and domain-specific tasks. The results emphasize the importance of task-specific alignment strategies, as performance gains in humanities and single-step reasoning were counterbalanced by declines in STEM and multi-step problems.

These findings underscore the need for adaptable alignment methodologies that cater to varied application requirements without compromising the core reasoning capabilities of LLMs. Future work could explore dynamic alignment approaches that adjust to contextual demands, offering insights into optimizing model utility across diverse real-world scenarios. This research lays a foundational framework for improving LLM alignment strategies, contributing to the broader goal of making AI systems both accessible and robust in their reasoning.

## 10 Future Work

- **Dynamic Alignment Approaches:** Investigate methods for dynamically adjusting alignment strategies based on task complexity and domain requirements. Adaptive alignment could prioritize clarity for simpler tasks while maintaining reasoning depth for complex, multi-step problems.
- **Addressing Alignment Trade-Offs:** Explore hybrid approaches that blend Direct Preference Optimization (DPO)[14] with reinforcement learning or other fine-tuning techniques to mitigate the trade-offs between reasoning depth and clarity.
- **Alignment for Specialized Applications:** Develop alignment strategies tailored to high-stakes domains such as healthcare, finance, or law. These strategies should emphasize interpretability, ethical reasoning, and transparency to meet domain-specific requirements.

## References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] https://arxiv.org/abs/2110.14168

[3] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. arXiv:1907.09190 [cs.CL] https://arxiv.org/abs/1907.09190

[4] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. arXiv:2212.07919 [cs.CL] https://arxiv.org/abs/2212.07919

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] https://arxiv.org/abs/2009.03300

[6] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina

Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] https://arxiv.org/abs/2305.18290

[8] Stephen Robertson. 2004. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation - J DOC* 60 (10 2004), 503–520. https://doi.org/10.1108/00220410410560582

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[10] Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. 2024. A Critical Evaluation of AI Feedback for Aligning Large Language Models. arXiv:2402.12366 [cs.LG] https://arxiv.org/abs/2402.12366

[11] Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023. Making Large Language Models Better Reasoners with Alignment. arXiv:2309.02144 [cs.CL] https://arxiv.org/abs/2309.02144

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

[13] Yotam Wolf, Noam Wies, Dorin Shteyman, Binyamin Rothberg, Yoav Levine, and Amnon Shashua. 2024. Tradeoffs Between Alignment and Helpfulness in Language Models with Representation Engineering. arXiv:2401.16332 [cs.CL] https://arxiv.org/abs/2401.16332

[14] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. arXiv:2404.10719 [cs.CL] https://arxiv.org/abs/2404.10719

[15] Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, and Bo Du. 2024.