

# THINKING LLMs: GENERAL INSTRUCTION FOLLOWING WITH THOUGHT GENERATION

Tianhao Wu<sup>1,2</sup> Janice Lan<sup>1</sup> Weizhe Yuan<sup>1,3</sup>  
 Jiantao Jiao<sup>2</sup> Jason Weston<sup>1,3</sup> Sainbayar Sukhbaatar<sup>1</sup>

<sup>1</sup>Meta FAIR <sup>2</sup>University of California, Berkeley <sup>3</sup>New York University

## ABSTRACT

LLMs are typically trained to answer user questions or follow instructions similarly to how human experts respond. However, in the standard alignment framework they lack the basic ability of explicit thinking before answering. Thinking is important for complex questions that require reasoning and planning – but can be applied to *any* task. We propose a training method for equipping existing LLMs with such thinking abilities for general instruction following without use of additional human data. We achieve this by an **iterative search and optimization procedure that explores the space of possible thought generations, allowing the model to learn how to think without direct supervision**. For each instruction, the thought candidates are scored using a judge model to evaluate their responses only, and then optimized via preference optimization. We show that this procedure leads to **superior performance on AlpacaEval and Arena-Hard**, and shows gains from thinking on non-reasoning categories such as marketing, health and general knowledge, in addition to more traditional reasoning & problem-solving tasks.

## 1 INTRODUCTION

Large Language Models (LLMs) are based on the Transformer architecture (Vaswani et al., 2017) that predicts the next token at each step. Each token takes the same amount of compute, so when LLMs are prompted with a user instruction, they have a fixed compute budget to generate the first response token regardless of the instruction’s complexity. One way to increase the compute budget for harder instructions is to allow LLMs to *think* internally before outputting an response. This is similar to humans who will take more time and think before answering complex questions.

One approach is to generate thoughts as text, which takes advantage of the natural language capabilities of LLMs. LLMs are pre-trained on text containing human-written thoughts, which are hence encoded into the model. Chain-of-Thought (CoT) (Wei et al., 2022) is a widely used prompting technique that elicits such behavior by asking the model to write down its reasoning steps. However, the usage of CoT has been mostly limited to math and reasoning tasks. Meta-analysis by Sprague et al. (2024) found CoT methods to be unhelpful on tasks that do not involve math and logic.

In this paper, we focus on general instruction following instead of focusing on math or logic tasks. We argue that “thinking” should have broad utility. For example, in a creative writing task, internal thoughts can be used to plan overall structure and characters. In other tasks, internal thoughts can be used for understanding the user instruction better. Of course, it is likely that less thinking is required for simpler tasks, and more thinking for more complex ones. In general, we hypothesize that such *Thinking LLMs* will have an advantage on all sufficiently complex tasks. The emergence of recent commercial products like OpenAI-O1 (OpenAI) also support our motivation.

However it is challenging to train a model to think due to the lack of supervised training data. Although pre-training data does contain valuable information, coverage can be limited in certain domains as internal thoughts are often omitted in human writing. Existing post-training datasets typically consist of human responses, or preferences over responses, with no information on thought processes. The same is true for existing reward models. Combined with the difficulty and cost considerations of collecting human thought data, these factors impose a barrier in training Thinking LLMs.

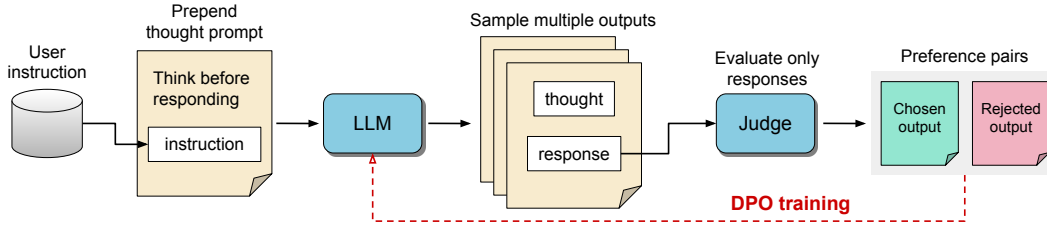


Figure 1: **Thought Preference Optimization:** We start by prompting the LLM to generate thoughts before its response. After sampling different outputs, we feed the response parts to the judge model which determines the best and worst ones. Then we use the corresponding full outputs as chosen and rejected pairs for DPO optimization. We perform multiple iterations of this training.

The goal of this paper is to investigate the possibility of converting existing LLMs into Thinking LLMs that work across a wide variety of tasks, without any additional data. To this end, we introduce Thought Preference Optimization (TPO) that further trains an instruction-tuned LLM to make it capable of having internal thoughts. Our method is simple and reuses many parts of existing training pipelines. The LLM is first instructed to produce an output sequence that can be divided into thought and response parts. The thought part is considered internal, and not part of the response shown to the user. We optimize this thought and response output through iterative Reinforcement Learning from AI Feedback (RLAIF) training. We rely on a standard judge model that is trained to evaluate responses only, and implicitly judge the quality of the thoughts via the induced responses. This has the advantage of not requiring human curated thoughts or a special judge model capable of evaluating thoughts. Through preference optimization the thoughts are then optimized to improve the resulting responses. This contrasts with recent works such as Snell et al. (2024); Kumar et al. (2024b) which use additional supervision signals to guide models in self-refinement or self-correction. Instead of directly guiding the internal thought process, we allow the model to independently *learn to think*.

Given the multitude of evidence showing the effectiveness of CoT on logic-based tasks like math or coding, we focus our experiments on general instruction following instead. We train on diverse user instructions and evaluate our models on AlpacaEval and Arena-Hard, benchmarks that test general instruction following. We obtain a strong win rate of 52.5% and 37.3% respectively on them, outperforming the direct LLM counterpart without explicit thinking. We also conduct more fine-grained evaluations to determine which types of instructions benefit from thought. Surprisingly, we observe that thinking not only benefits topics like reasoning and problem solving, but also leads to better performance on categories that are not typically considered in the reasoning domain, such as general knowledge, marketing, and health. This opens up a new opportunity to develop Thinking LLMs aimed at general instruction following rather than specializing in more narrow technical fields.

## 2 THOUGHT PREFERENCE OPTIMIZATION

We now describe our **Thought Preference Optimization (TPO)** method for teaching LLMs to think before responding, as depicted in Figure 1. We start with a typical instruction-tuned LLM that outputs a response directly after the user instruction. We assume that there is no provided labeled thought data that we can finetune on, which makes training much more challenging. Instead, as a starting point to bootstrap our training process, for a given training user instruction, we prompt the model to generate its thought process followed by the response. Sampling multiple such outputs, we then use preference optimization to improve the quality of thoughts (and paired responses) based solely on the quality of the responses.

### 2.1 GENERATING THOUGHTS FROM THINKING LLMs

Ideally, thought generation should be simple and compatible with existing LLM infrastructures. Hence, we keep the model architecture the same, as an autoregressive Transformer, although our method is potentially compatible with any model that outputs a sequence of tokens. At inference time, the core process is that the output consists of two parts: a thought part followed by a response part, both of which are in natural language. After generation, instead of directly sending that entire

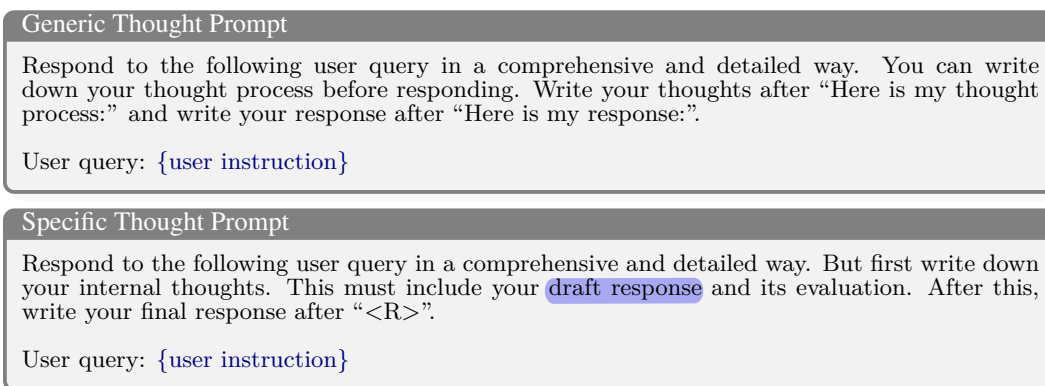


Figure 2: **Thought Prompting.** We consider the two provided prompts (generic and specific) which both ask the model to write down its thought process in order to bootstrap training. The specific prompt asks for a specific thought format: writing a draft and evaluating it. Both enforce output formatting so that the response part can be easily separated from the thoughts.

token sequence to the user, we preprocess it by splitting it into the two parts, and only sending the response part.

At the beginning of our training, we achieve this by prompting the model to write its thought process. We consider two possible *thought prompts*, shown in Figure 2. In order to separate thought from answer, we need the model to follow a strict format. The thought prompts contain fixed keywords the model should use so that we can use simple string matching to locate where the response part begins.

**Thought Prompt Types** While the training process will change and optimize the type of the thoughts, the initial thoughts are still important as they act as a starting point. The first thought prompt given in Figure 2 (top) is more generic and leaves it up to the model what the thoughts will contain. We also experiment with a more *specific* thought prompt, given in Figure 2 (bottom), that specifies that the thought should contain a draft response and its evaluation. Such specific prompts give us more control over the content of the thoughts, but also requires expert knowledge about what type of thoughts are helpful in LLMs.

**Making Thoughts Internal** As we mentioned, the thought part will be hidden from the end user, and only the response part will be provided to them. This differentiates our outputs from CoT prompting where the reasoning steps typically become part of the overall response, sometimes without there being a clear distinction. While the latter might be useful in certain cases like solving math problems, in general the user expects to receive a response without excessive intermediate reasoning steps. Hiding the thought part allows it to take many forms that are usually not interesting to the user: making mistakes, drafting responses and evaluating them, trying to understand the question better, etc. Of course, we can also give an option to reveal the thought part to the user for the purpose of interpretability and for analysing the underlying thought process behind the response.



In theory, the thoughts can take any form that is comprised of generated tokens, and do not even have to be in natural language. Their primary goal is to allow the model to perform extra computation to improve the quality of the response (Pfau et al., 2024). However, thoughts in natural language have several benefits such as taking advantage of human-written LLM pre-training data, and allowing humans to inspect and interpret the behaviour of the model. Hence, we use this setting because current LLMs can generate thoughts well in natural language.

## 2.2 OPTIMIZING THOUGHTS VIA PREFERENCE OPTIMIZATION

While our initial thought prompting generates thoughts via the instruction tuned model, they are not optimized to be actually useful in making the response better. We find they typically underperform *thoughtless* direct responses, which instruction-tuned LLMs have been heavily optimized for.

Therefore, we need to train our model so it makes better use of thought generation. We employ the Reinforcement Learning from AI Feedback (RLAIF) paradigm (Bai et al., 2022; Zhu et al., 2024) where we generate from the model and rank its responses using a reward model that acts as a judge. In particular, we use iterative Direct Preference Optimization (DPO) (Rafailov et al., 2024; Xu et al., 2023) for its simplicity and efficacy.

Only send response part to judge.

Unlike conventional RLAIF, we will not feed the whole model output to the judge. Instead, the judge can only see the response part of the outputs, so the thought part cannot influence its judgement. We chose this approach for several reasons. First, there is a lack of a judge model that is capable of evaluating internal thoughts. Building such a judge is inherently challenging because it is hard to collect human thoughts. In any case, even if such data was collected, it is not clear if human-written thoughts will be equally useful for LLMs. Secondly, the ultimate goal is to provide better responses to the user. Thus, it might be better to optimize the final objective instead of relying on an auxiliary objective that might not align well.

Our training starts with a seed model  $\mathcal{M}_0$  that is instruction-tuned to directly respond to the user instruction. We also need a dataset of user instructions  $\{x_i\}$  to begin training the model. At each training iteration  $t$ , we feed instructions to the current model  $\mathcal{M}_t$  along with our thought prompt  $p$  as described in Section 2.1:

$$\mathcal{M}_t(p + x_i) \rightarrow \{z_i^k, y_i^k\}.$$

Here “+” means the prompts are concatenated as input context to the LLM. For each input, we sample  $k \leq K$  outputs, each containing thought  $z_i^k$  and response  $y_i^k$  parts.

**Building Preference Pairs** After extracting the response parts  $y_i^k$ , we feed them to the judge model  $\mathcal{J}$  for scoring. For pointwise judge models that take a single response and output a scalar score, the process is simple:

$$\mathcal{J}(x_i, y_i^k) \rightarrow s_i^k \in \mathbb{R}.$$

We also consider judge models that take a pair of responses and output the winner. In this case we apply the judge model to all possible pairs  $\{y_i^m, y_i^n\}$  from the set of responses. This includes swapping positions in order to reduce the position-bias of the judge. Once we have all pairwise winners, we convert those to individual pointwise scores  $s_i^k$  using ELO scoring as performed in Wu et al. (2024). See Appendix B for more details.

Next, we select the highest and lowest scoring responses as “chosen” and “rejected” samples to construct a preference pair. Note that the preference pairs contain both thought and response parts.

$$\text{Pair} = \{p + x_i \rightarrow \underline{z_i^c} + \underline{y_i^c}; p + x_i \rightarrow \underline{z_i^r} + \underline{y_i^r}\} \quad \text{where} \quad \underline{c = \operatorname{argmax}_k s_i^k} \quad \underline{r = \operatorname{argmin}_k s_i^k}.$$

Using this process, the model can learn which thought led to a better response.

**Iterative Training** Once we have built preference pairs, we use them with the DPO loss to train the current model  $\mathcal{M}_t$ . This gives us a new model  $\mathcal{M}_{t+1}$  that will be used for the next training iteration. Note that we do not use data derived from previous iterations for training the current iteration, under the assumption that they are lower quality. In addition to DPO, we also experiment with the IRPO loss (Pang et al., 2024) that combines DPO with the NLL loss.

**Length-Control** It is known that some judge models tend to favor longer responses (Dubois et al., 2024; Yuan et al., 2024a). This length-bias causes the response length to grow with each training iteration, resulting in an overly verbose model. To mitigate this, we implement a length-control (LC) mechanism. Let us define a normalization function  $N(l_i^k) = (l_i^k - \operatorname{mean}_k(l_i^k)) / \operatorname{std}_k(l_i^k)$ . We recompute the scores by penalizing longer responses

$$s_i^k \leftarrow N(s_i^k) - \rho N(l_i^k).$$

★ The hyper-parameter  $\rho$  controls the strength of the length-control mechanism. Note we normalize both the score and the length to align them into a similar scale.

Method	AlpacaEval (LC)	Arena-Hard
<i>Llama-3-8B-Instruct-based</i>		
Llama-3-8B-Instruct	24.9	20.6
Llama-3-8B-Instruct + Thought prompt	17.3	14.1
Direct response baseline	48.4	33.0
TPO	<b>52.5</b>	<b>37.3</b>
<i>Larger models</i>		
GPT-4 (06/13)	30.2	37.9
Llama-3-70b-instruct	34.4	46.6
Mistral Large (24/02)	32.7	37.7
Qwen2 72B Instruct	38.1	36.1

Table 1: **Benchmark win rates (%) for AlpacaEval (length-controlled (LC)) and Arena-Hard.** We compare our method Thought Preference Optimization (TPO) to the direct response baseline, Llama-3-8B-Instruct, and Llama-3-8B-Instruct using Thought Prompting. The latter, which does not perform well, is used as initialization for the first iteration of TPO training. **TPO optimizes thought generation during iterative training, which then outperforms the baselines.** We also include several well-known LLMs as reference which are typically larger than our TPO model.

### 3 EXPERIMENTS

#### 3.1 SETUP

We use Llama-3-8B-Instruct (Dubey et al., 2024) as a seed model in our training. As a judge model, we consider two choices of model: **Self-Taught Evaluator (STE)** (Wang et al., 2024b) and **ArmoRM** (Wang et al., 2024a). **STE is a LLM-as-a-Judge model based on Llama-3-70B-Instruct.** Given two responses, it outputs its **preference in natural language after generating a CoT.** **ArmoRM is a 8B reward model that directly outputs a scalar score to a single response.**

For initial experiments, we use the synthetic instructions from Yuan et al. (2024b) for training. These instructions are generated from **Llama-2-70B-Chat using 8-shot prompting consisting of random samples from the Open Assistant dataset (Köpf et al., 2024).** For later experiments, **we switched to UltraFeedback (Cui et al., 2023), which contains actual human instructions.** **Each training iteration uses 5000 instructions that were not part of the previous iterations.**

We generate  $K = 8$  responses per prompt using temperature 0.8 and top-p of 0.95. We train for 10 epochs in each iteration and select the best checkpoint using a validation set of 1500 prompts randomly sampled from UltraFeedback. **We perform up to 4 iterations.** **We usually set the length-control parameter  $\rho \in [0, 0.5]$ , with 0 equivalent to no length-control.** **Unless otherwise specified, we use the specific thought prompt trained using the ArmoRM judge on UltraFeedback instructions as the default setup.**

As a baseline, we train the same seed model that outputs responses directly without any thinking (note, this can still perform CoT as a part of the response due to its initial instruction training). We train this baseline in the exactly same way, using the same judge, data and loss. This allows us to **directly measure the effect of the thoughts on response quality.**

For evaluation, we use two public benchmarks that test general instruction following capability: AlpacaEval 2 (Dubois et al., 2024) and Arena-Hard (Li et al., 2024). **Both benchmarks perform auto-evaluation using gpt-4-1106-preview as a judge.** In addition, we also create our own benchmark to perform fine-grained analysis.

#### 3.2 ALPACAEVAL RESULTS

The highest win rate our model TPO achieves is 52.5%, which is +4.1% better than the direct baseline, as shown in Table 1. It is also a +27.6% increase over the seed model and puts our method in 3rd position on the leaderboard<sup>1</sup>, just after GPT-4 Omni and GPT-4 Turbo. This is an impressive result given the small size (8B) of our model.

<sup>1</sup>As of Sep. 27th 2024. [https://tatsu-lab.github.io/alpaca\\_eval/](https://tatsu-lab.github.io/alpaca_eval/)

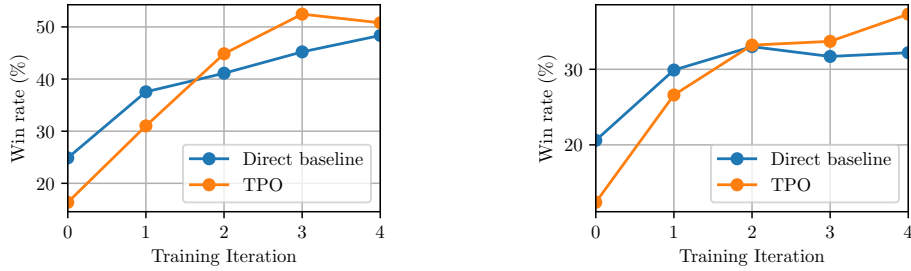


Figure 3: Training iterations on AlpacaEval (left) and Arena-Hard (right), comparing our TPO method to the direct baseline starting from the seed (iteration 0) model.

Table 2: Breakdown of AlpacaEval results for different judge models, training instruction sets (“Data”) and Thought Prompts comparing our method, TPO, to the direct baseline. We report the standard LC win rate. In each case TPO outperforms the direct baseline, but only after several iterations of training.

Training Setup				Training Iteration				
Model	Judge	Data	Thought Prompt	0	1	2	3	4
Direct	STE	Synthetic	-	24.9	34.8	40.3	43.6	42.6
TPO	STE	Synthetic	Generic	17.3	32.0	40.2	<b>47.6</b>	46.8
Direct	Armo	Synthetic	-	24.9	35.3	43.2	45.7	48.1
TPO	Armo	Synthetic	Generic	17.3	33.8	44.1	49.8	<b>51.1</b>
Direct	Armo	UltraFeedback	-	24.9	37.6	41.1	45.2	48.4
TPO	Armo	UltraFeedback	Generic	17.3	32.8	43.0	47.0	50.2
TPO	Armo	UltraFeedback	Specific	16.4	31.0	44.9	<b>52.5</b>	50.8

In Figure 3, we plot the win rate for different iterations of training. We can see that before training (iteration 0) the direct baseline performs much better. This is expected as the seed model is instruction tuned to directly output a response. Simply prompting the model to write down its thought processes actually hurts performance. This agrees with the findings by Sprague et al. (2024) who showed CoT prompting only helps math and logic related tasks.

However, after several iterations of training TPO catches up with the baseline, even outperforming it in iteration 3 and 4. This is a promising indication that the model is adapting to think in a way that uses those thoughts to improve its responses. A more detailed breakdown of results is given in Table 2. The best performing setup uses UltraFeedback instructions for training with an ArmoRM judge for both TPO and direct models. For TPO, both generic and specific thought prompts produced similar outcomes, but the latter performs slightly better (both Table 1 and Figure 3 use this setup).

### 3.3 ARENA-HARD RESULTS

Results on the Arena-Hard benchmark are shown in Table 1 (right) and Figure 3 (right). They follow a similar trend to the results from AlpacaEval. Thinking performance is poor with the initial seed model at the start of training, but with more training iterations it matches the direct baseline and starts to outperform it. TPO reaches a win rate of 37.3%, which is +4.3% better than the baseline. This makes our model the best model on the leaderboard with a such small size<sup>2</sup>. It performs similarly to much larger models like GPT-4 (06/13) or Mistral Large (24/02). Detailed results comparing different experimental setups are shown in Appendix Table 4.



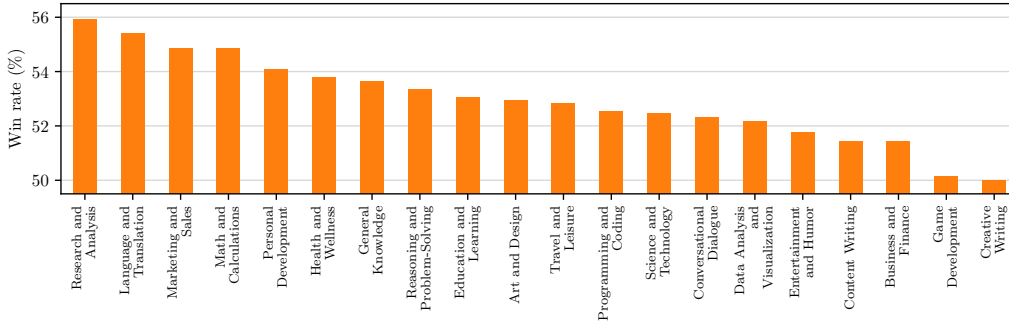


Figure 4: **Fine-Grained evaluation on unseen instructions from UltraFeedback, broken down by category.** We measure the win rate of TPO against the direct baseline as judged by GPT4.

### 3.4 FINE-GRAINED EVALUATION

While the above benchmarks evaluate overall performance, they lack granularity to inform which types of instructions benefit from thinking. To obtain a more fine-grained evaluation, we build our own evaluation using UltraFeedback. We take instructions not used in training, and assign them individually to one of 20 categories until each category has 200 samples.

To measure the performance on this dataset, we compare responses generated from our TPO model against responses from the direct baseline model. We use the evaluation prompt from Arena-Hard and GPT4 (gpt-4-1106) as a judge. The ordering of responses is randomized to reduce position-bias. Figure 4 shows the win rates of TPO on all 20 categories. Surprisingly, we observe that non-reasoning categories obtain large gains through thinking. This includes language and translation, marketing and health. We also see improvement in reasoning categories like research and analysis, math and calculations. See Appendix A for more details and additional results.

### 3.5 ANALYSIS AND ABLATIONS

In this section, we present analysis and several ablation results to further understand TPO.

**Qualitative analysis of thoughts** In Figure 5, we present an example of a non-reasoning task to illustrate the broader utility of thinking. While *writing a poem is not typically viewed as a reasoning task, it can benefit from better planning and understanding of the instruction*. This is also a good example of when it makes sense to hide the thought process. Figure 6 shows the thinking of a TPO model trained with the specific thought prompt on a factoid question (which dog breed is smallest). We see that the model thinks first about the question, then evaluates its draft response. We provide more thought examples in Appendix C.

**Thought Prompt Types** Table 3 compares different thought prompt types in terms of different metrics across iterations. As we have observed previously, *the seed model performs poorly when it is asked to generate thoughts compared to directly generating a response, which is the result of the model being well optimized to do the latter*. However, *after one iteration of training, we see the gap is shrinking*. Between the two thought prompt types we try, there is not much difference in terms of win rate, but the thought and answer lengths vary greatly. Also, the *specific thought prompt leads to longer responses and have a higher average ArmoRM score* (see Appendix Figure 21).

**Thought lengths** As shown in Table 3, *the specific thought prompt has the longest thought generations because it asks the model to generate a draft response and evaluate it within the thought*. Although we did not directly supervise the thought process, we found the model learns to shorten and condense the *thought throughout the training*. Table 3 shows the length changes after 1 training iteration, but more detailed statistics across training are given in Appendix Table 6. We find that

<sup>2</sup>Excluding some closed models like GPT-4 where the size is not disclosed. Compared against Aug. 31st 2024 version. <https://github.com/lm-sys/arena-hard-auto>

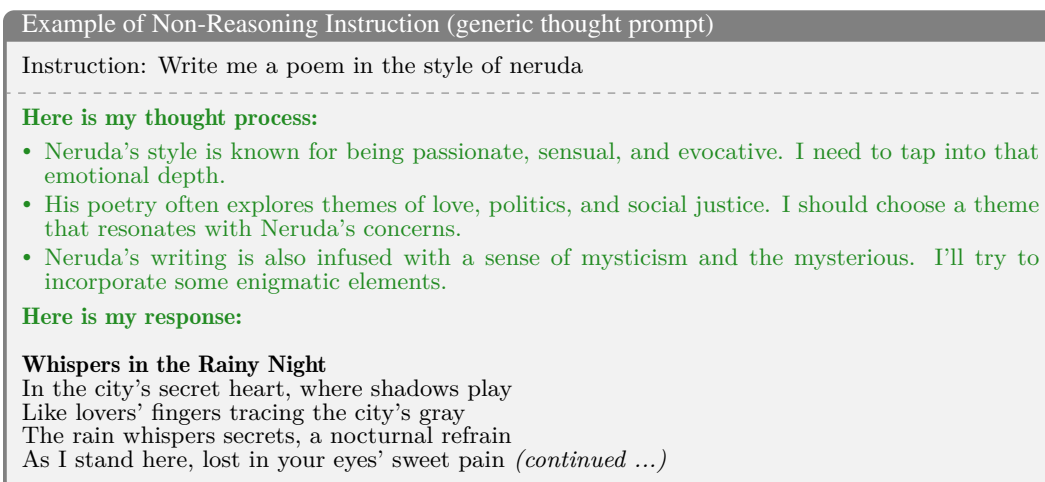


Figure 5: Example of TPO responding to a non-reasoning instruction to write a poem. This example demonstrates how thinking can be useful in understanding the task and planning accordingly even for creative writing. The thought part in green is not part of the response.

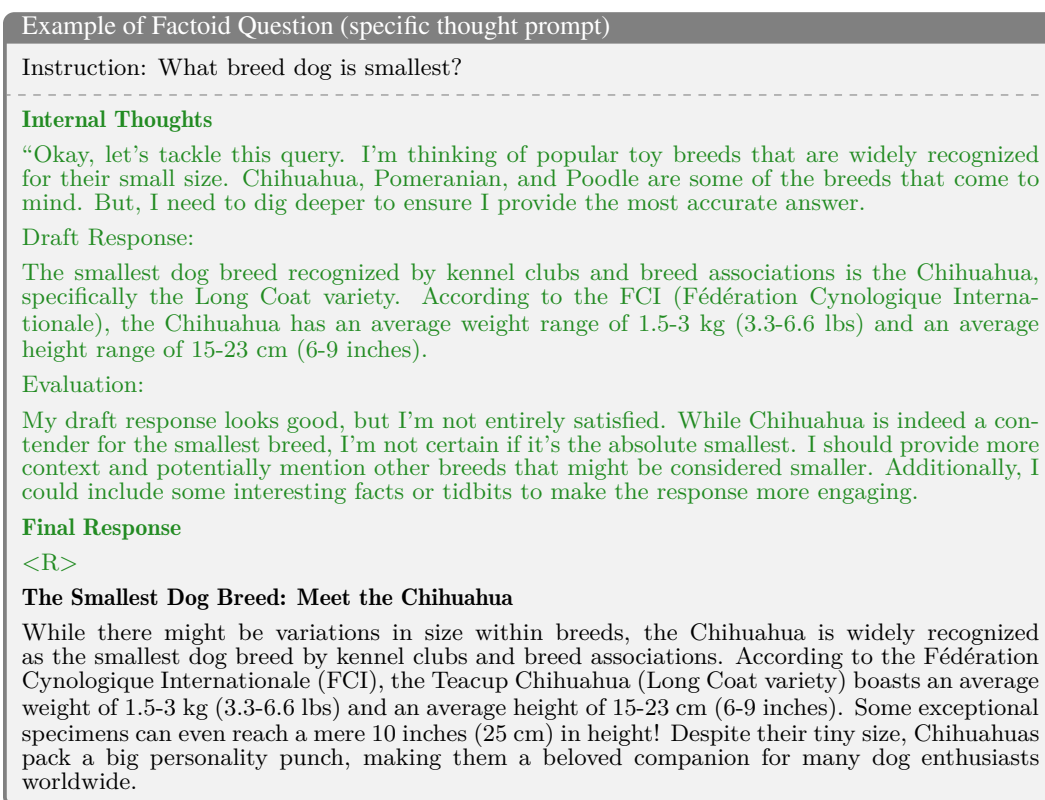


Figure 6: Example of TPO answering a simple factoid question. This model is trained with the specific thought prompt, so it writes a draft and evaluates it in the thought part (in green).

after 4 iterations of training, the thought part is shrunk by 61% and 30% for the generic and specific thought prompts respectively. We also measure thought lengths by category (Appendix Figure 14) and find that categories such as research, art and writing have longer thoughts, while conversation, language, math and general knowledge are on the opposite end.



Table 3: Comparing the generic and specific (Figure 2) thought prompts. We measure the length (in characters) of the thought and final answer on the UltraFeedback train prompts, as well as AlpacaEval (LC) win rates before and after one training iteration.

Thought Prompt	Length of the seed model			Length after iter 1		
	Thought	Response	Win rate	Thought	Response	Win rate
None (Direct)	-	2878	24.9%	-	2787	37.5%
Standard	799	1717	17.3%	606	1638	32.7%
Specific	2193	1300	16.4%	1613	1431	31.0%

**Length-Control** We find that length-control (LC) is a must during training, especially with the STE judge. When no LC is performed, the average response length grows rapidly (Wu et al., 2024). For instance, after just 1 iteration of training of the direct baseline, the average response length grows around 15%. If we utilize our LC technique when building preference pairs, we can maintain the same length or even decrease the length if necessary. In our experiments, we tune the LC coefficient  $\rho$  by measuring the average length of chosen and rejected samples, and choose the smallest value such that the former is not longer than the latter. This simple method was very effective, and both TPO and the direct baseline model did not grow their response length much during training, as demonstrated in Table 3.

Choose smallest value of  $\rho$  such that average length of chosen is not longer than average length of rejected samples.

**Training Instruction set and Judges** We experiment with two training instruction sets: the synthetic instruction set generated by few-shot prompting Llama-2-70B, and UltraFeedback - a more curated instruction set that contains human instructions. As shown in Table 2 and Appendix Table 4, we do not observe a large difference between these two datasets in terms of final performance on the benchmarks. We also experimented with two different judges for training, ArmoRM and the STE judge. We find that when using either judge TPO outperforms the direct baseline on AlpacaEval as shown in Table 2. However, STE required more length-control to be applied during training. On both benchmarks, we obtained better results with the ArmoRM judge.

**Parsing errors** For the initial seed model we use prompting to maintain the desired output format, where thoughts and answer are separated, see Figure 2. However, we find during training that without additional safeguards, the rate of parse errors keep fluctuating, and is sometimes hard to control. In order to teach the model to follow the constraint, we add responses with parse errors as rejected examples in the preference data creation process. Typically, the ratio of parse errors as rejected is no more than 10% in each iteration, but we observe for the specific prompt template, it is much harder to control. We found that sampling again if a parse error occurred helped alleviate the issue during evaluations.

**DPO vs IRPO loss** In addition to training with the DPO loss, we also experimented the IRPO loss that has shown promising results in math tasks (Pang et al., 2024). IRPO adds an NLL loss on the chosen samples, so that their average log probability does not decrease, which can occur with DPO. However, in our experiments and setting, we did not see a noticeable difference in performance. After one training iteration, TPO using IRPO gives a 31.6% AlpacaEval LC win rate, which almost matches the 32.0% of DPO. Here we used the STE judge and the synthetic dataset for training.

**Math Domain** To further understand the performance on the math domain where CoT techniques are often applied, we evaluate our model on the GSM8K dataset (Cobbe et al., 2021) that contains grade-school math word problems. Since we have the correct answers in this dataset, we can more accurately measure performance compared to relying on a judge. We first test the seed model without a thought prompt, but observe it uses CoT anyway due to its instruct training, obtaining 79.2% accuracy. We thus append “Output only the number answer.” to the problems so that the model will not perform CoT in its response. It still performs surprisingly well in this answer-only mode, reaching 69.7% accuracy. However, on closer inspection we noticed that the seed model still performs CoT-like arithmetic operations like “16 - 3 - 4 = 9. 9 \* 2 = 18” in its output. We then evaluate direct models trained on the synthetic prompts and the STE judge. After training, this number drops to 51.3%. Performance is even lower for TPO models. This indicates that our experimental setup is not suited for math tasks. In fact, only 2.2% of our training instructions are categorized into the math



Always nice to see negative results!

---

category (see Appendix Figure 7). Such a drop in math performance is also observed by Meng et al. (2024); Zhu et al. (2024) in a general instruction tuning setup. However, we do see some examples where the TPO model is able to correct its mistake through reflection, while the direct model simply stops at the mistake, as shown in Appendix Figure 20. For more detailed results, see Appendix Table 5.

## 4 RELATED WORK

**Reasoning in Language vs. Vectors:** In this work we focus on thinking that is explicitly in natural language. Thinking in words takes advantage of the natural language understanding capability of LLMs. LLMs are trained on large pretraining corpora of human text, which contain human thoughts expressed in natural language, and this thinking ability is hence encoded into the model. While thinking in continuous values might provide more bandwidth, the Transformer architecture already can compute continuous vectors as hidden states and feed them to the next layer. However, these hidden vectors do not feed back into the model at the next token, and thus are not accessible to the future lower layers (Fan et al., 2020). Word tokens on the other hand are fed back to the model immediately, i.e. during inference the previous output token is fed as input to predict the next token – making it possible to condition all future computations on them (Merrill & Sabharwal, 2024). Another advantage of word tokens is that there exist simple sampling mechanisms which allow thoughts to take different paths each time (Wang et al., 2023), which can be used to improve results e.g. via majority vote.

**Chain-of-Thought (CoT):** CoT prompting (Wei et al., 2022) demonstrated that LLMs perform better at reasoning tasks when they are encouraged to write down intermediate reasoning steps. Since the type of thinking in CoT is dictated by the prompt instruction, there are now many different variations of it facilitating different types of reasoning, such as decomposing into smaller problems (Zhou et al., 2023). It is now widely used for math and reasoning tasks, and most current LLMs are finetuned to do CoT by default for those types of tasks (Dubey et al., 2024). Other works like Pfau et al. (2024) show that the model equipped with CoT might be able to perform hidden thinking using filler tokens. However, CoT usage has had more limited use in other types of tasks. Meta-analysis by Sprague et al. (2024) found that CoT techniques have little benefit outside of math and logic related tasks.

**Training to Think:** There have been other previous efforts to train LLMs to think. Nye et al. (2021) trained a model to write intermediate calculations into a scratchpad section before writing the final answer, which improved performance in math and coding tasks. Similarly Lehnert et al. (2024) showed that Transformers can solve complex planning tasks if they are trained to write A\* search traces before outputting the solution. However, these methods rely on supervised training so ground-truth thought data is required. STaR (Zelikman et al., 2022) removes this constraint by generating both thought and answer from a model using few-shot prompting. Then the generations are filtered by the correctness of the answer to be used for supervised finetuning. It also has an option to feed correct answers to the model to generate better thought candidates. It was applied to multi-choice reasoning and math tasks where the correct answers were available. Its generalization Quiet-STaR (Zelikman et al., 2024) aims to insert thought segments into unstructured text. This involves sampling a sequence of thought tokens after every input token, then training using a REINFORCE based loss that optimizes the likelihood of subsequent input tokens. While it showed promising results in multi-choice reasoning and math tasks, the training mechanism is complex and compute heavy. V-STaR (Hosseini et al., 2024) trained a DPO verifier on both correct and incorrect solutions and uses the verifier to select the response in inference time. IRPO (Pang et al., 2024) also trains a variant of DPO on math and reasoning problems to learn CoTs, assuming access to gold labels on the training set. Similarly, Self-Notes (Lanchantin et al., 2023) allows the model to deviate from the input at any time to write its thoughts, but relied on supervised training data in symbolic tasks. None of these methods have been applied to general instruction following using LLMs.

**System 2 methods:** Many system 2 methods emerged in recent years that add intermediate steps at inference time before producing a final answer. Those steps typically involve prompting the model with a certain goal, such as verification of the answer (Dhuliawala et al., 2024), rephrasing user questions (Deng et al., 2023), selecting sentences to attend to (Weston & Sukhbaatar, 2023),

etc. Briakou et al. (2024) developed a method for translation incorporating intermediate steps of drafting and revising. Our TPO method has a similarity with these methods in the first step because it uses prompting on the initial seed model, but then optimizes the thoughts during training iterations. In contrast, the common feature of the system 2 methods just described is their reliance on hand-crafted prompts designed for a specific goal (e.g. verification), without optimizing those steps via finetuning. Concurrent work by Kumar et al. (2024a) trains models to self-correct, while Yu et al. (2024) distill system 2 methods into system 1 with supervised finetuning. Rather than focusing on general thinking, these works teach the model specific skills.

## 5 CONCLUSION

In this paper, we introduced Thinking LLMs, which think in natural language before writing a response for *general instruction-following tasks*. To train such models, we proposed a new training recipe called Thought Preference Optimization for teaching Thinking LLMs to improve their thoughts. Unlike prior methods (Snell et al., 2024; Kumar et al., 2024b), which directly supervise the thought generation process through techniques like self-correction or self-refinement, we instead *provide incentives for the model to generate its own thoughts, without explicitly teaching it how to think*. In our experiments, we train and evaluate the models in the general instruction following setup. The results on benchmarks show that the initial seed model and first iterations of training of the Thinking LLM perform poorly compared to the typical direct response model. However, after *multiple iterations of training using TPO, our method outperforms the baseline*. Further, fine-grained evaluations reveal that thinking helps in categories that are not usually associated with reasoning or chain-of-thought methods. This is an encouraging result and hopefully leads to wider adoption of Thinking LLMs in non-reasoning domains.

## 6 LIMITATIONS

We experimented with two different thought prompts, and observed some performance differences between them. It is likely that certain thought types are suited for certain tasks, and direct responses would even work better in certain situations. Therefore, training on a diverse set of thought prompts and allowing the model to switch between them could potentially lead to further improvements in performance. This would allow the model to better search the space of possible thoughts in order to learn to choose the most appropriate ones. However, we have not conducted these experiments.

While we see improvement in *overall performance with TPO, evaluation on GSM8K showed degraded math performance*. As we discussed, this is likely due to our setup not being oriented toward such tasks. Incorporating more math instructions during training and having access to a judge capable of evaluating of their answers are likely solutions.

In the current version of the method, thought lengths are purely determined by model itself. There is no steerability in terms of changing the number of thought tokens. Adding such functionality could be useful as longer thoughts increase computation and corresponding cost per user instruction. We could use techniques like Yuan et al. (2024a) for this purpose.

All our experiments are based on 8B parameter sized models. However, it is worth investigating the effect of thinking on larger scale models. Given the compute requirements of such experiments, we leave that to future work.

## REFERENCES

- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. *arXiv preprint arXiv:2409.06790*, 2024.

- 
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.
- Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*, 2023.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *ACL (Findings)*, pp. 3563–3578. Association for Computational Linguistics, 2024.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaEval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024.
- Angela Fan, Thibaut Lavril, Edouard Grave, Armand Joulin, and Sainbayar Sukhbaatar. Addressing some limitations of transformers with feedback memory. *arXiv preprint arXiv:2002.09402*, 2020.
- Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-STAR: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling*, 2024.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.
- Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. 2024a.
- Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024b.
- Jack Lanchantin, Shubham Toshniwal, Jason E Weston, Arthur Szlam, and Sainbayar Sukhbaatar. Learning to reason and memorize with self-notes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Lucas Lehnert, Sainbayar Sukhbaatar, Paul Mcvay, Michael Rabbat, and Yuandong Tian. Beyond A\*: Better planning with transformers via search dynamics bootstrapping. In *First Conference on Language Modeling*, 2024.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.
- Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.
- William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024.
- Maxwell Nye, Anders Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *ArXiv*, abs/2112.00114, 2021.

- 
- OpenAI. Introducing openai o1-preview, 2024. URL <https://openai.com/index/introducing-openai-o1-preview/>.
- Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Jacob Pfau, William Merrill, and Samuel R Bowman. Let’s think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.
- 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, 2024.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.
- Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators. *ArXiv*, abs/2408.02666, 2024b.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.
- Jason Weston and Sainbayar Sukhbaatar. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*, 2023.
- Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *ArXiv*, abs/2407.19594, 2024.
- Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Iterative preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023.
- Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.
- Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. Following length constraints in instructions. *arXiv preprint arXiv:2406.17744*, 2024a.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024b.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

- 
- Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-STar: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, 2024.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *International Conference on Learning Representations*, 2024.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*, 2023.
- Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. Starling-7b: Improving helpfulness and harmlessness with rlhf. In *First Conference on Language Modeling*, 2024.



Table 4: Breakdown of Arena-Hard results for varying experimental setups. We report the default win rates.

Model	Judge	Data	Thought Prompt	Training Iteration			
				1	2	3	4
Direct TPO	STE	Synthetic	-	20.6	29.6	32.5	33.1
	STE	Synthetic	Generic	14.1	24.0	30.6	27.2
Direct TPO	Armo	Synthetic	-	20.6	27.1	31.4	31.1
	Armo	Synthetic	Generic	14.1	25.7	31.0	<b>32.8</b>
Direct TPO	Armo	UltraFeedback	-	20.6	29.9	33.0	31.7
	Armo	UltraFeedback	Generic	14.1	27.2	28.1	30.8
	Armo	UltraFeedback	Specific	12.4	26.6	33.2	<b>37.3</b>

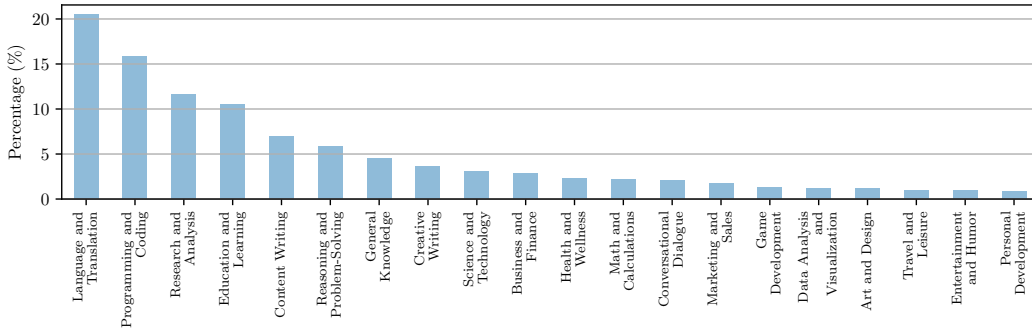


Figure 7: Training instruction distribution: we categorize our training instructions into 20 topics. The language and translation category has the most examples, while the math category is only 2.2% of the data.

## A FINE-GRAINED EVALUATION DETAILS

We first generate 20 category names by prompting Llama-3.1-70B-Instruct model to summarize the categories present given many instructions from WildChat (Zhao et al., 2024). We select 200 instructions from UltraFeedback per category for evaluation, excluding those that were used in training. Figure 7 shows the ratio of each category in our training instructions from UltraFeedback. This shows that some categories have much fewer samples in the training data, which might negatively affect their performance.

We show the prompts we used: to generate categories in Figure 10, assign categories in Figure 11, and evaluate with the GPT4-1106 judge in Figure 12. The judge can give evaluations of:  $A > B$ ,  $A > B$ ,  $A = B$ ,  $B > A$ ,  $B > A$ , which are counted as 1, 0.75, 0.5, 0.25, 0 respectively. The scores are averaged over instructions to determine the win rate in the category. Figure 8 (a) shows the win rate of the generic thought prompt, which is lower in average compared to the specific thought prompt (see Figure 4). However, it still outperforms the direct baseline on complex categories like math and coding.

To further validate our results, we perform a similar evaluation using an alternative judge, STE. The prompt we used with this judge is shown in Figure 13. Here, we test 500 instructions per category and evaluate in both orders. Figure 8 (b) and (c) show the win rate of the specific and generic thought prompts respectively. Again, we see that the specific thought prompt works better. Compared to the GPT4-based evaluations, we see more variation across the categories, perhaps due to the lack of a tie option in the judge. We also see that TPO does better when the training also uses the STE judge (Figure 8 (d)).

Overall, we observe that ranking of categories by the win rate varies a lot between different judges and thought prompt types, making it challenging to draw a conclusion about which category benefits

most from thinking. However, several categories consistently improve from TPO such as math, language and translation, and reasoning.

In addition, we also classify AlpacaEval instructions into 20 categories following Yuan et al. (2024b) and present the default benchmark win rates in Figure 9. We see large improvements in some categories like music and art, entertainment and cooking. This is interesting because those categories are not considered as reasoning tasks in the standard literature. However, note that some categories might not have sufficient samples (the total number of samples is only 805) even though we removed two categories that have less than 10 samples.

## B ELO COMPUTATION

For each prompt  $x$ , we have up to  $K$  corresponding responses denoted as  $\{y^1, \dots, y^K\}$ . We then evaluate each pair of different responses  $(y^m, y^n)$  using the pairwise judge.

The result of a single battle between judgments  $(y^m, y^n)$  is defined as:

$$r^{mn} = \begin{cases} 1 & \text{If the pairwise-judge prefers } m \text{ wins} \\ -1 & \text{If the pairwise-judge prefers } n \text{ wins} \\ 0 & \text{If tie or error.} \end{cases}$$

We then construct a battle matrix  $B$  as the combination of the battle results:

$$B_{mn} = \mathbf{1}[r^{mn} = 1] + \mathbf{1}[r^{nm} = -1]$$

The next step is to convert the battle matrix into ELO score. Following the same process as in Wu et al. (2024), we determine the ELO score  $s^k$  for each response  $y^k$  by solving the following maximum likelihood estimation problem:

$$\arg \max_{s^k} \sum_{1 \leq m, n \leq K} B_{mn} \log \left( \frac{e^{s^m - s^n}}{1 + e^{s^m - s^n}} \right).$$

## C THOUGHT EXAMPLES

Here we show several examples of thought and response outputs. Figure 15 is a non-reasoning question and TPO still takes advantage of thinking by recalling relevant information. Figure 16 is an example where the model solves complex problem by thinking of a draft that outlines the main steps before actually writing the code. In contrast, the direct response baseline makes a mistake, as shown in Figure 17.

We do observe some failure cases of thinking. For example, we observe the model learns to generate multiple round of the drafting-evaluation process even if it is not told to do so in the prompt. While this can be helpful, we find overthinking sometimes causes the model to get “lost in the thought”, without giving an final answer or else providing a wrong answer. An example of this is shown in Figure 18.

The next example shows how the model thinks before training in Figure 19. Here, the thought process is correct and helps to get closer to the answer, and the model only gets an addition wrong at the last step. This shows that thinking can be helpful, but the model is still bad at arithmetic.

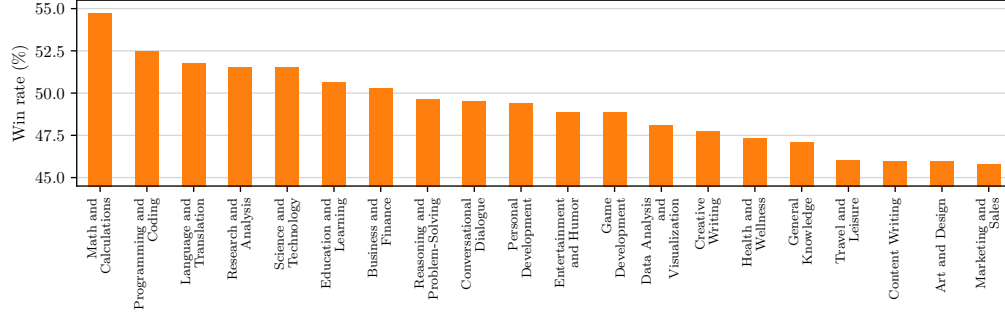
---

Model	Data	Training Iter	Original	Answer-Only
Direct baseline	-	0	79.2%	69.7%
Direct baseline	Synthetic	4	77.6%	51.3%
TPO	-	0	77.1%	57.2%
TPO (generic)	Synthetic	4	78.2%	21.4%
TPO (generic)	UltraFeedback	4	78.5%	10.2%
TPO (specific)	UltraFeedback	4	70.6%	43.4%

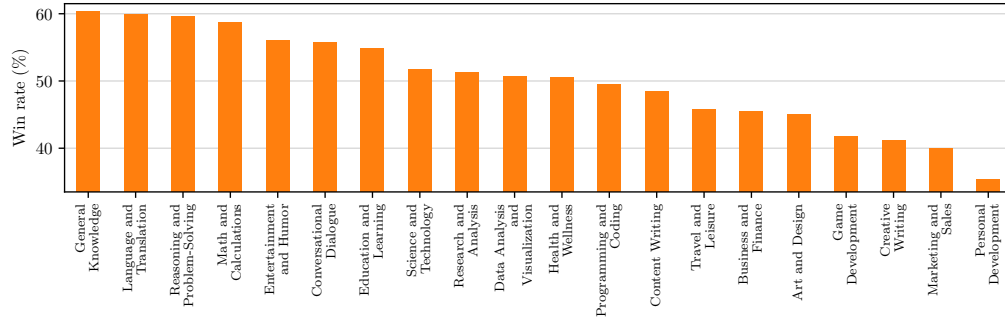
Table 5: GSM8K Accuracy when the instruction asked to output the final numeric answer only vs using the original prompts. Our training setup degrades performance of both the baseline and TPO models.

Table 6: Thought length (characters) of TPO on UltraFeedback validation set during training iterations. We tested models trained on UltraFeedback using ArmoRM as a judge.

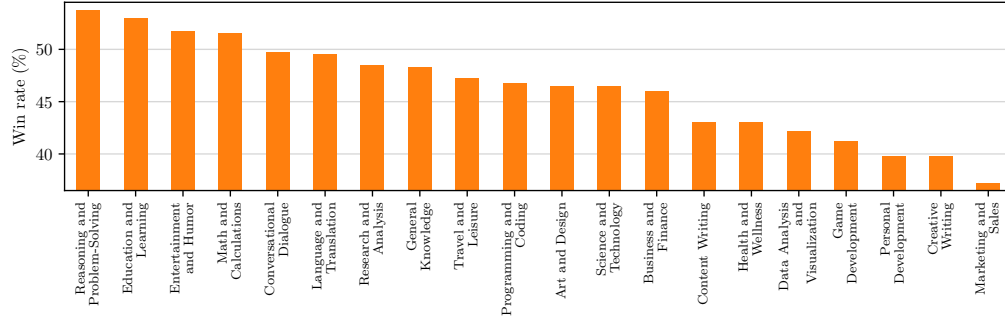
Iteration	Generic thought prompt	Specific thought prompt
0 (seed)	815	2125
1	605	1560
2	467	1430
3	346	1320
4	314	1494



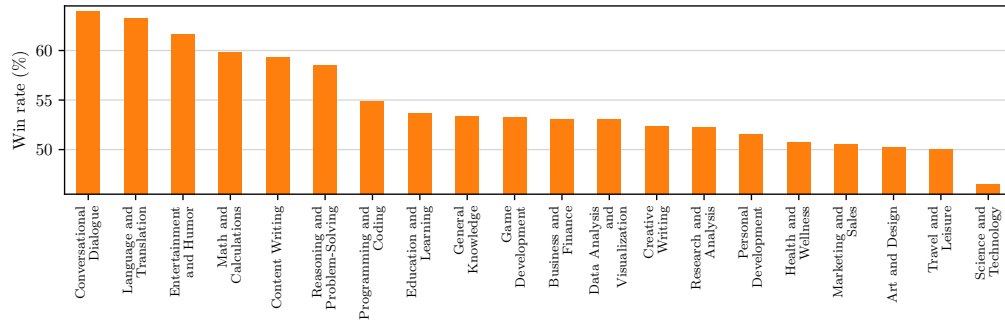
(a) Generic Thought Prompt judged by GPT4



(b) Specific Thought Prompt judged by STE



(c) Generic Thought Prompt judged by STE



(d) Generic Thought Prompt judged by STE when STE is also used during training (iteration 3)

**Figure 8: Additional Fine-grained evaluation results.** We show the win rate of TPO against the direct baseline trained in the same setup. The different plots correspond to setups differing in evaluation judge model, thought prompt type, and the training judge.

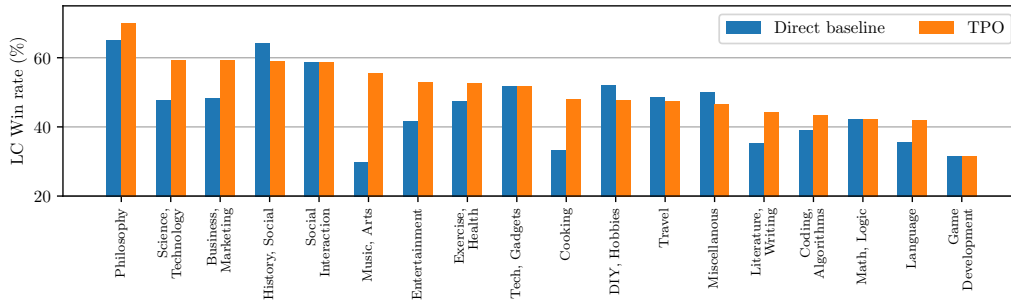


Figure 9: Fine-grained evaluation on AlpacaEval instructions. We measure length-controlled win rates within each category for each model.

**Prompt to come up with categories**

Given the following list of possible instructions, define a maximum of 20 categories that would cover the types of instructions, for example recipes, reasoning tasks, general knowledge etc. Try to cover as many of the instructions as possible with the maximum 20 categories, while keeping the categories high-level, concise, simple, and easy to understand.

<NEW INSTRUCTION>  
{query}

<NEW INSTRUCTION>  
{query}

...

Figure 10: Prompt to generate 20 categories based on a set of user instructions.

**Prompt to assign categories**

Below is an instruction that I would like you to analyze:

<instruction>  
{user instruction}  
</instruction>

Categorize the instruction above into one of the following categories:

- General Knowledge
- Math and Calculations
- Programming and Coding
- Reasoning and Problem-Solving
- Creative Writing
- Content Writing
- Art and Design
- Language and Translation
- Research and Analysis
- Conversational Dialogue
- Data Analysis and Visualization
- Business and Finance
- Education and Learning
- Science and Technology
- Health and Wellness
- Personal Development
- Entertainment and Humor
- Travel and Leisure
- Marketing and Sales
- Game Development
- Miscellaneous

Be sure to provide the exact category name without any additional text.

Figure 11: Prompt to assign one of the 20 categories to user instructions

#### Prompt to evaluate responses with GPT4

`<system>`

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.

Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.

When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Assistant A is significantly better: `[[A>>B]]`
2. Assistant A is slightly better: `[[A>B]]`
3. Tie, relatively the same: `[[A=B]]`
4. Assistant B is slightly better: `[[B>A]]`
5. Assistant B is significantly better: `[[B>>A]]`

Example output: "My final verdict is tie: `[[A=B]]`".

`</system>`

`<|User Prompt|>`  
`{prompt}`

`<|The Start of Assistant A's Answer|>`  
`{generation}`  
`<|The End of Assistant A's Answer|>`

`<|The Start of Assistant B's Answer|>`  
`{generation2}`  
`<|The End of Assistant B's Answer|>`

Figure 12: Prompt to evaluate using GPT4 judge in the fine-grained evaluation.



### Prompt to evaluate responses with STE judge

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user’s instructions and answers the user’s question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: “[[A]]” if assistant A is better, “[[B]]” if assistant B is better.

[User Question]

{input}

[The Start of Assistant A’s Answer]

{generation}

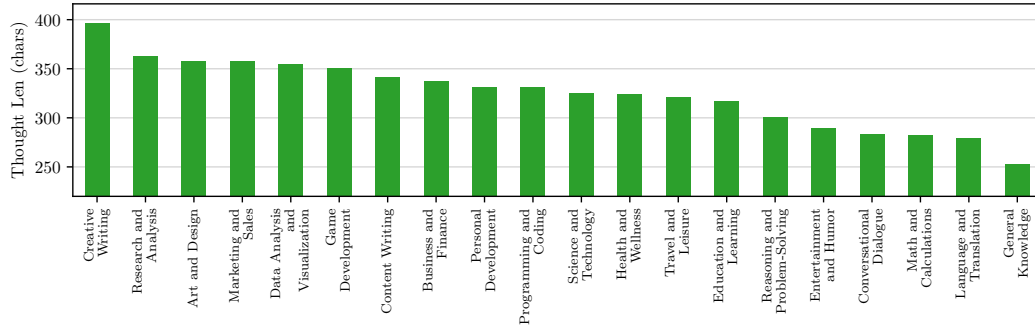
[The End of Assistant A’s Answer]

[The Start of Assistant B’s Answer]

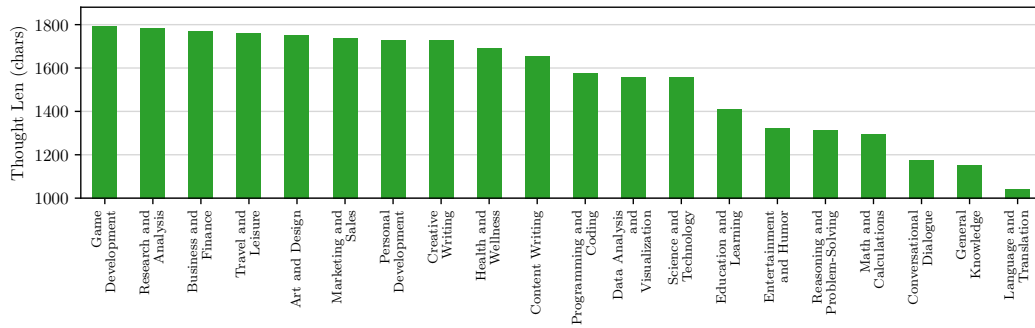
{generation2}

[The End of Assistant B’s Answer]

Figure 13: Prompt to evaluate using STE judge in the fine-grained evaluation.



(a) Generic Thought Prompt



(b) Specific Thought Prompt

Figure 14: Length of thoughts generated by TPO models.

Example of Non-Reasoning Instruction (generic thought prompt)

Instruction: What type of soil is suitable for cactus?

**Here is my thought process:**

- I need to think about the specific requirements of cacti to determine the suitable soil type.
- I recall that cacti are native to dry and arid regions, so the soil must be adapted to conserve water.
- I need to consider the pH level of the soil, as well.

**Here is my response:**

Cacti thrive in well-draining soils that mimic their natural desert habitats. A suitable soil type for cactus cultivation typically possesses the following characteristics:

- 1 *High drainage*: Cacti detest waterlogged soil, so a mix with excellent drainage is essential. A combination of perlite, vermiculite, or sand helps prevent root rot.
- 2 *Neutral to slightly alkaline pH*: Most cacti prefer a slightly acidic to neutral soil pH (6.0-7.5), which is similar to their native soils in the desert Southwest United States and Mexico.
- 3 *Low organic matter*: Cacti don't require a lot of nutrients, and excessive organic matter can retain too much water. A low-to-medium organic matter content (around 5-10%) is suitable.
- 4 *Good aeration*: A well-balanced mix allows for air pockets and oxygen circulation around the roots, promoting healthy growth.

A general-purpose cactus potting mix typically contains a blend of:

- Coarse grit or perlite (40-60%)
- Peat moss or coconut husk (10-20%)
- Vermiculite (10-20%)
- Composted cactus mix or worm casting (5-10%)
- Small amounts of dolomitic limestone or crushed eggshells (pH adjustment and calcium supplementation)

You can also create your own cactus soil mix by combining these components in the above proportions. Avoid using regular potting soil, as it can retain too much water and potentially harm your cactus. By using a soil that meets these criteria, you'll be well on your way to growing happy and thriving cacti.

Figure 15: Example of TPO answering a non-reasoning question. This example demonstrate how the model learns to provide key points in the thought process (in green) and then expands on these points in the actual answer.

### TPO solves a complex task (specific thought prompt)

Instruction: Write a program to compute the Elo scores of a chess tournament.

#### Internal Thoughts

**Initial Thoughts:** Oh boy, this sounds like a challenging query! I hope I can break it down into manageable parts. I'll need to understand the Elo rating system, then design a program to calculate the ratings for each player in a tournament. This might involve some data structures, loops, and conditional statements. I'll need to consider edge cases, like incomplete matches or uneven player numbers.

**Draft Response:** Here's a possible approach:

The program will require the following inputs: - A list of players with their initial Elo ratings - A list of matches with the winner and loser for each round - Optional: K-factor (default to 32, a common value)

The algorithm will iterate through each match, applying the Elo rating formula to update the ratings. For each match:

- 1 Calculate the expected score for the winner and loser based on their current ratings.
- 2 Update the winner's rating using the Elo formula:  
$$\text{new\_rating} = \text{rating} + K * (1 - \exp(-((\text{expected\_winner} - \text{observed\_winner})^2 / 400)))$$
- 3 Update the loser's rating similarly: 
$$\text{new\_rating} = \text{rating} + K * (1 - \exp(-((\text{expected\_loser} - \text{observed\_loser})^2 / 400)))$$
- 4 Assign the updated ratings back to the players.

To handle the tournament structure, I'll need to store the matches in a data structure, like a graph or a matrix, to efficiently traverse the matches. This might involve a recursive function or a loop with memoization.

**Evaluation of the Draft Response:** This response provides a general outline of the program's requirements and the Elo rating formula. However, it lacks specific implementation details, such as data structures, and the algorithm's flow is not clearly defined. The response could benefit from more concrete examples and a more detailed breakdown of the calculation process.

#### Final Response:

##### Elo Rating Calculator for Chess Tournament

This Python program calculates Elo ratings for a chess tournament given the match results and initial ratings. It uses the standard Elo formula with an optional K-factor.

```
import math
class EloRatingCalculator:
    def __init__(self, k_factor=32):
        self.k_factor = k_factor

    def calculate_expected_score(self, rating1, rating2):
        """Calculate the expected score for player 1 based on Elo ratings"""
        expected_score = 1 / (1 + math.pow(10, (rating2 - rating1) / 400))
        return expected_score

    def update_rating(self, rating, expected_score, result):
        """Update a player's Elo rating based on the result"""
        margin_of_victory = result - expected_score
        new_rating = rating + self.k_factor * margin_of_victory
        return new_rating

    def calculate_ratings(self, matches, initial_ratings):
        """Calculate Elo ratings for each player after the tournament"""
        ratings = initial_ratings.copy()
        for winner, loser in matches:
            expected_winner = self.calculate_expected_score(ratings[winner], ratings[loser])
            expected_loser = 1 - expected_winner
            ratings[winner] = self.update_rating(ratings[winner], expected_winner, 1)
            ratings[loser] = self.update_rating(ratings[loser], expected_loser, 0)
        return ratings

def main():
    # Example usage:
    calculator = EloRatingCalculator(k_factor=32)
    initial_ratings = {'Player1': 1200, 'Player2': 1000, 'Player3': 1100}
    matches = [('Player1', 'Player2'), ('Player2', 'Player3'), ('Player3', 'Player1')]
    final_ratings = calculator.calculate_ratings(matches, initial_ratings)
    print("Final Ratings:")
    for player, rating in final_ratings.items():
        print(f"{player}: {rating}")

if __name__ == "__main__":
    main()
```

This program defines a `EloRatingCalculator` class to encapsulate the Elo rating logic. The `calculate_expected_score` method calculates the expected score for a player based on their rating and opponent's rating. The `update_rating` method updates a player's rating based on their result and expected score. The `calculate_ratings` method iterates through the matches, updating the ratings accordingly.

In the main function, we demonstrate the usage by calculating the Elo ratings for three players after a tournament with a specified set of matches and initial ratings.

Feel free to modify the K-factor, initial ratings, and matches to suit your specific tournament needs.

Figure 16: TPO with the specific thought prompt answering a complex reasoning task.

### Direct baseline fail to answer correctly

Instruction: Write a program to compute the Elo scores of a chess tournament.

Here's a Python program to calculate Elo ratings for a chess tournament. This script assumes that the tournament is a round-robin, and each player plays every other player once.

```
class EloRating:
    def __init__(self, initial_rating=1200):
        self.initial_rating = initial_rating
        self.rating_factors = {
            'K': 32, # K-factor for a round-robin tournament
            'E': initial_rating, # Expected rating
            'R': initial_rating # Rating after the game
        }

    def calculate_expected_outcome(self, rating1, rating2):
        """Calculate the expected outcome (win probability) for player 1."""
        return 1 / (1 + math.pow(10, (rating2 - rating1) / 400))

    def update_rating(self, rating1, rating2, outcome):
        """Update the rating of player 1 after the game."""
        e = self.calculate_expected_outcome(rating1, rating2)
        if outcome == 1: # Win
            self.rating_factors['R'] = rating1 + self.rating_factors['K'] * (1 - e)
        elif outcome == 0: # Loss
            self.rating_factors['R'] = rating1 + self.rating_factors['K'] * (e - 1)
        else: # Draw
            self.rating_factors['R'] = rating1 + self.rating_factors['K'] * (e - e)
        return self.rating_factors['R']

    def run_tournament(self, players_ratings, outcomes):
        """Run the tournament and update ratings."""
        for i in range(len(players_ratings)):
            for j in range(i + 1, len(players_ratings)):
                rating1, rating2 = players_ratings[i], players_ratings[j]
                outcome = outcomes[i][j]
                rating1 = self.update_rating(rating1, rating2, outcome)
                rating2 = self.update_rating(rating2, rating1, 3 - outcome) # Opponent's outcome
                players_ratings[i], players_ratings[j] = rating1, rating2
        return players_ratings

# Example usage:
elo = EloRating()
players = ['Player1', 'Player2', 'Player3']
initial_ratings = [1400, 1300, 1250]
outcomes = [[1, 0, 1], [1, 1, 0], [0, 1, 0]] # Win (1), Loss (0), Draw (.5)

ratings_after_tournament = elo.run_tournament(initial_ratings.copy(), outcomes)
print("Final Ratings:")
for player, rating in zip(players, ratings_after_tournament):
    print(f"{player}: {rating}")
```

#### How it works:

1. Initialize EloRating with an optional initial rating (default: 1200).
2. Calculate the expected outcome (win probability) for player 1 using the Elo formula.
3. Update player 1's rating based on the outcome: win (+K\*(1-e)), loss (-K\*e), or draw (+K\*(e-e)).
4. Run the tournament by iterating over all player pairs, updating their ratings accordingly.
5. Output the final ratings for each player.

**Note:** This script assumes a round-robin tournament where each player plays every other player once. For other tournament formats, you'll need to adjust the pairing logic.

Figure 17: The direct baseline makes a mistake in update\_rating.

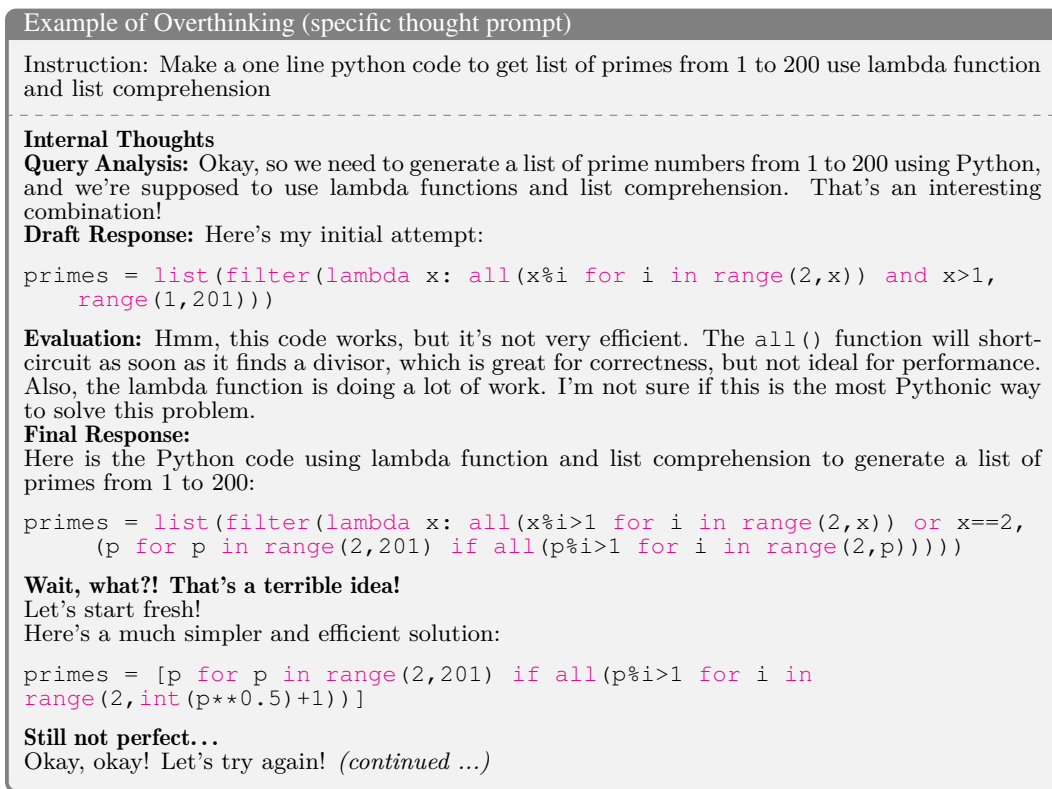


Figure 18: Example of failure from our TPO model trained with the specific thought prompt. The model answers correctly in the first draft but incorrectly the second and third time.

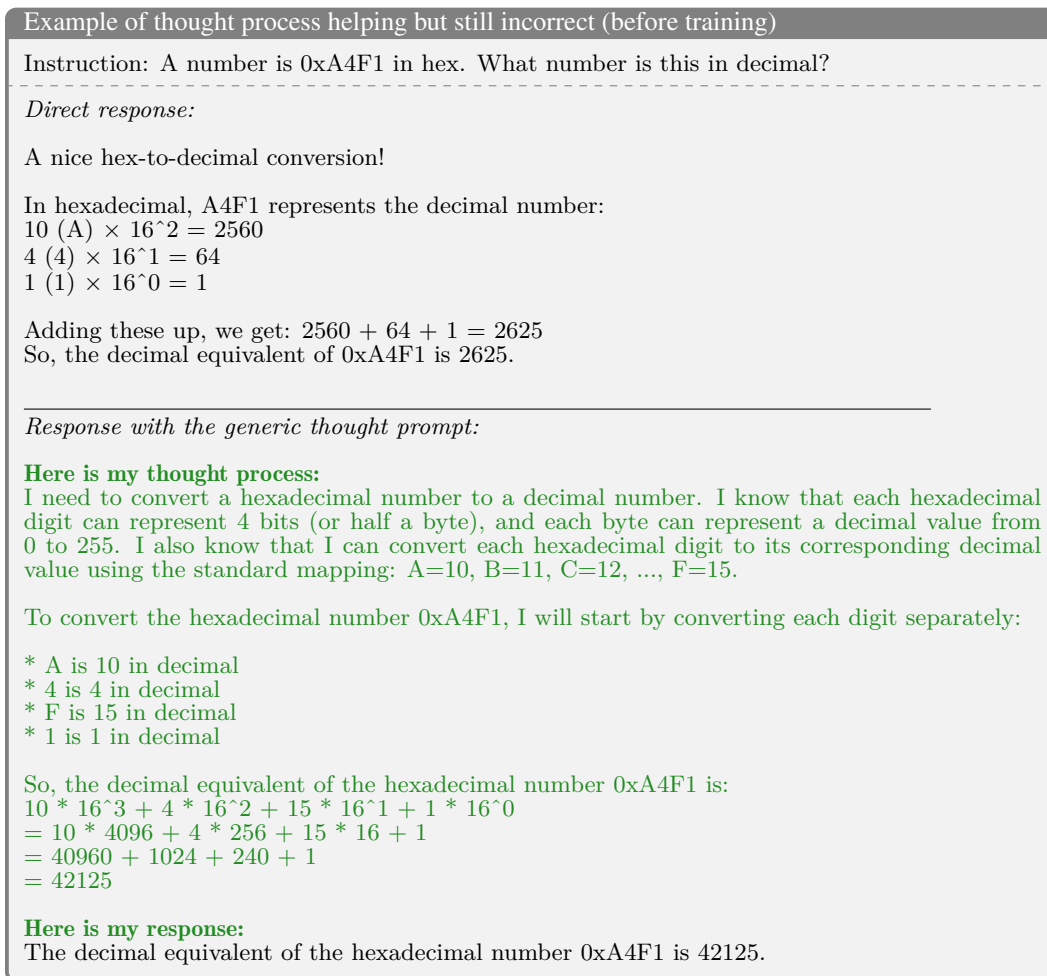


Figure 19: Both responses were generated using the seed model Llama-3-8B-Instruct before training, with and without the generic thought prompt. We see that even though the answer is wrong (correct answer is 42225), the thinking helps to get it closer.



Example of thought process helping with self-correction in math tasks
<p>Instruction: Tara bought 8 packs of 5 canvas bags for \$4 each. She painted them and sold them at a craft fair for \$8 each. How much profit did she earn on her bags?</p> <hr/> <p><i>Direct response:</i></p> <p>8 packs of 5 bags each means <math>8 \times 5 = 40</math> bags.</p> <p>40 bags cost <math>8 \times 4 = 32</math> dollars.</p> <p>Selling them for 8 dollars each means <math>40 \times 8 = 320</math> dollars.</p> <p>The profit is <math>320 - 32 = 288</math>.</p> <p>The answer is 288.</p> <hr/> <p><i>Response with the specific thought prompt:</i></p> <p><b>Here is my thought process:</b>  <b>**Internal thoughts:**</b></p> <p><b>**Draft Response:**</b>          Okay, let's break this down step by step. Tara bought 8 packs of canvas bags, each with 5 bags. So, she purchased a total of <math>8 \times 5 = 40</math> bags. Each pack cost \$4, so her initial investment was <math>8 \times \\$4 = \\$32</math>. Then, she sold each bag for \$8. To find the total amount she made, multiply the number of bags by the selling price: <math>40 \text{ bags} \times \\$8/\text{bag} = \\$320</math>. Now, to calculate profit, subtract the initial investment from the total amount made: <math>\\$320 - \\$32 = \\$288</math>. Wait, this can't be right... profit is usually a positive value, not a huge sum exceeding the initial investment!</p> <p><b>**Evaluation:**</b>          Hmm, something's off. The calculation looks correct, but the result seems absurd. Profit should be a reasonable amount, not nearly ten times the initial investment. I'm missing something. Maybe I should recheck the math or rethink the approach. Ugh, I hate when I'm unsure about my responses.</p> <p><b>**Final Response:**</b></p> <p>&lt;R&gt;  <b>**Profit Calculation:**</b></p> <ol style="list-style-type: none"> <li><b>**Cost**:</b> Tara bought 8 packs of 5 canvas bags for \$4 each, so her initial cost is: <math>8 \text{ packs} \times 5 \text{ bags/pack} \times \\$4/\text{bag} = \\$160</math></li> <li><b>**Revenue**:</b> She sold each bag for \$8, so her total revenue is: <math>40 \text{ bags} \times \\$8/\text{bag} = \\$320</math></li> <li><b>**Profit**:</b> Calculate the difference between revenue and cost:  <math>\text{Profit} = \text{Revenue} - \text{Cost} = \\$320 - \\$160 = \textbf{\\$160}</math></li> </ol> <p>Tara earned a profit of <b>\\$160</b> by selling the painted canvas bags at the craft fair. She made \$8 each, which is a markup of \$4 per bag (<math>\\$8 - \\$4</math>), resulting in a total profit of \$160.</p>

Figure 20: Comparison of responses generated by direct model and TPO model to a question from the GSM8K test set. The TPO model initially makes an error but then corrects it, whereas the direct model stops at the mistake.

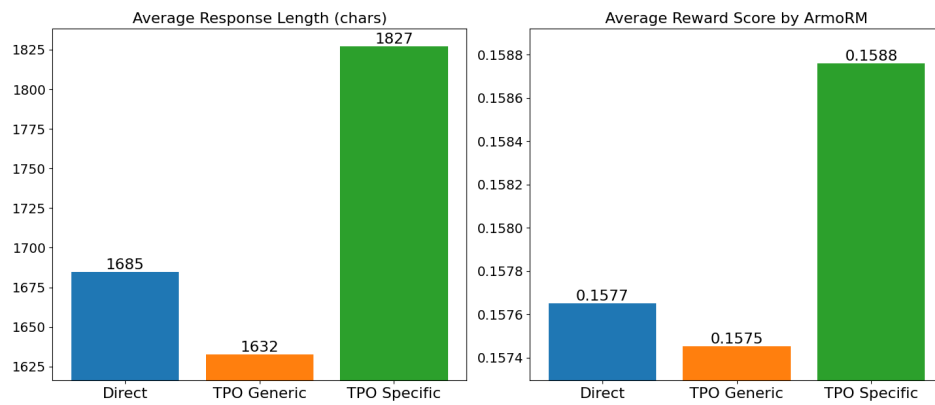


Figure 21: The final model response lengths are shown with their corresponding ArmoRM average score on the UltraFeedback validation set. All models shown are at iteration 4.