Compositional preference models for alignment with scalable oversight

Dongyoung Go[†*] Tomek Korbak[‡] German Kruszewski[§]
Jos Rozen[§] Marc Dymetman[+]

[†]NAVER Corp [*]Yonsei University [‡]University of Sussex [§]Naver Labs Europe [+]Independent Researcher

Paper URL QR code

ABSTRACT

Can we combine LM capabilities with human insight to make PMs more interpretable and reliable?

- We show that decomposing one global preference assessment into several interpretable features is more robust to overoptimization and more aligned with the preference

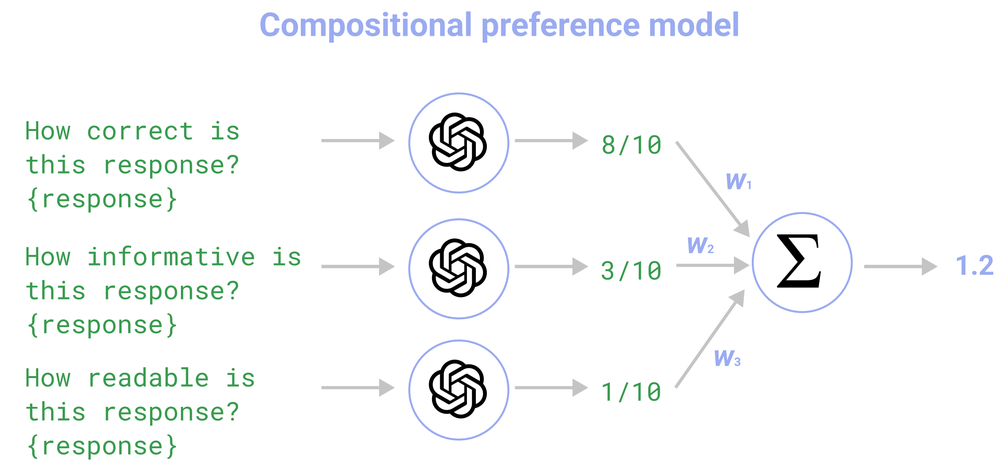
The key contributions :

- Introducing CPM, a novel framework for learning PMs that is **more robust to overoptimization**
- Allowing for **more transparent supervision** and **effective preference alignment**, by decomposing the preference problem into a series of intuitive features linked to human preferences, and employing an LLM as a feature score extractor.
- Systematically investigating the performance of CPMs on a diverse array of dimensions, including model robustness, generalization, robustness to overoptimization, and effectiveness for preference alignment

Summary of our method.

- Step 1: Feature decomposition. We decompose a hard question (e.g. “is this text preferable?”) into a series of easier questions (e.g. “is this text informative?”, “is this text readable?”) that are easier to evaluate for an LM and easier to inspect for a human overseer.
- Step 2: Feature scoring. Then, a prompted LM (e.g. ChatGPT) with pre-specified prompt templates assigns a numerical value to each feature. (e.g. informativeness: 3/10, readability: 1/10)
- Step 3: Aggregation. Finally, the feature scores are combined into a single preference score using a logistic regression classifier trained to predict human preference judgements (i.e. which of two texts a human would prefer).

Compositional Preference Model (CPM)



Prompt template used in Experiments

```
You will be given a conversation between a human and an AI
assistant.
Your job is to judge whether assistant's reply {attribute_desc}.
Score that on a scale from 1 to 10 where 1 means {attr_min} and 10
means {attr_max}. Here's the conversation:
Human:
{history}
Assistant:
{reply}
Answer by outputting a number from 1 to 10 (and nothing else).
Answer:
```

Experiments

Experimental Setup

- Dataset: **HH-RLHF** dataset, **SHP** dataset, sampled **20K** single-turn data points
- Features: 13 features: `helpfulness, specificity, intent, factuality, easy-to-understand, relevance, readability, enough-detail, biased, fail-to-consider-individual-preferences, repetitive, fail-to-consider-context and too-long`
- Best-of-n (BoN): A simple yet effective method that has been shown to be competitive with more advanced techniques such as reinforcement learning.
  We **generate n responses using an initial LM** and compare the robustness of two related PMs, by **choosing the sample with maximum PM score** and measuring the gap between their average scores.
  We generate up to 25,600 BoN responses, with 256 responses for each of 100 prompts in a held-out test set. We use Flan-T5-Large (780M) as the initial LM to generate the responses.

Experiment1_ Robustness to overoptimization

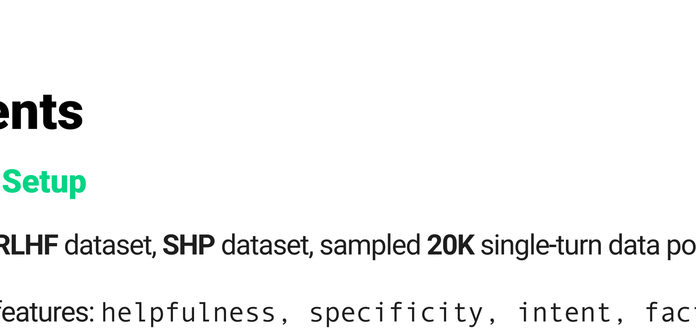


Figure: Overoptimization experiment. Dashed line means proxy PM used for BoN selection, corresponding solid line means god PM

- We construct a synthetic dataset where the output of one PM (defined to be the **“gold PM”**) is assumed to be the ground truth for human preferences.
- Here, the score of the **gold PM refers to the genuine preference**, by the experimental design
- CPMs show improved robustness to overoptimization than standard PMs. **The gap between gold and proxy PM scores is smaller for CPMs**, and the gold PM score begins to diverge later than for standard PMs, even though the same capabilities are used for both (Flan-T5-XL, 3B)

Experiment2_ Quality evaluation

Win Rate	HH-RLHF	SHP
CPM-GPT-3.5	**0.810** (.)	**0.672** (.)
CPM-Flan-T5	0.742 (0.034)	0.580 (0.045)
Standard PM	0.588 (0.030)	0.564 (0.037)

Table: Win rate over initial generation after BoN sampling based on each PM. Except CPM-GPT-3.5, we independently conduct 10 rounds of BoN(n=16) samplings and report the average win rate with standard error.

- The best-of-n samples obtained using CPMs are **consistently preferred** over samples obtained using conventional PMs, as **evaluated by Claude** as a human approximator. This suggests that CPMs are not only robust, but also more effective for learning about human preferences.

Results_ Model interpretability

CPM-GPT-3.5		CPM-Flan-T5	
Feature	Coefficient	Feature	Coefficient
helpfulness	0.246	fail-to-consider-context	0.420
enough-detail	0.235	enough-detail	0.244
factuality	0.187	factuality	0.227

Table: Three largest CPM coefficients on HH-RLHF dataset.

- The interpretability of our model is enhanced by the fact that the **feature coefficients** provide a direct indication of the factors that most influence the CPM’s decisions.
- This information can help understand the CPM’s internal workings.

Dialogue	Features				
	helpfulness	readability	enough-detail	factuality	fail-to-consider-context
Q: What things can I create with java programming?					
R1: Java is a programming language that is designed to create applications that run on a computer.	1.209	0.559	0.913	1.690	2.117
R2: Robots	0.509	0.559	-0.075	-0.569	-0.239
Q: Why is Lucifer referred to as the “morning star”?					
R1: Lucifer is the Morning Star because it is the brightest star in the sky.	1.209	0.880	1.901	0.843	0.232
R2: angel	-0.890	0.239	-0.569	-0.851	-0.239
Q: I’d like to evaluate some art perhaps. Can you tell me some history about Mona Lisa, the painting?					
R1: The Mona Lisa is a famous painting in the Louvre, Paris, France. It is attributed to Leonardo da Vinci.	1.909	0.880	1.901	1.690	2.588
R2: It was painted by Gian Lorenzo da Sica in the late 18th century. The name of the painting is inspired by the story of the famous Italian painter Leonardo da Vinci. The painting has been described as having a “poetic” quality.	0.859	0.239	1.901	0.278	-0.239

Table: Examples of feature values of CPM. Each feature value can provide information about which aspects of the response are good or bad.

- The features extracted by the LM enable intuitive explanation of generated responses.
- This **allows supervising complex behavior** in a human-interpretable way.
- By decomposing a hard preference (“This text is not preferable.”) into a series of easier features (“This text is generally unhelpful, as it is easy to read but has little detailed information”), it allows easier inspection for a human overseer.