Improving Preference Extraction In LLMs By Identifying Latent Knowledge Through Classifying Probes

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are often used as automated judges to evaluate text, but their effectiveness can be hindered by various unintentional biases. We propose using linear classifying probes, trained by leveraging differences between contrasting pairs of prompts, to directly access LLMs' latent knowledge and extract more accurate preferences. Through extensive experiments using models of varying size from four different families and six diverse datasets assessing text quality evaluation and common sense reasoning, we demonstrate that both supervised and unsupervised probing approaches consistently outperform traditional generation-based judgement while maintaining similar computational costs. These probes generalise under domain shifts and can even outperform finetuned evaluators with the same training data size. Our results suggest linear probing offers an accurate, robust and computationally efficient approach for LLM-as-judge tasks while providing interpretable insights into how models encode judgement-relevant knowledge. Our data and code will be openly released in the future.

1 Introduction

Chatbot Large Language Models (LLMs) are often trained using Reinforcement Learning with Human Feedback (RLHF) over preference datasets in order to increase honesty, helpfulness, and harmlessness (Christiano et al., 2017; Stiennon et al., 2020; Bai et al., 2022). This manifests as an increase in value/judgement alignment with humans, allowing for the use of such models as stand-in replacements for human raters on various tasks of evaluation (Zheng et al., 2023b; Shen et al., 2023; Zeng et al., 2023; Stephan et al., 2024; Zhong et al., 2022). This approach, commonly known as *LLM-as-a-Judge*, is particularly powerful for its fast and automatic nature.

In these tasks, LLM evaluators perform direct score-based assessment or pairwise comparisons, conventionally through generation-based prediction, where models are prompted to output numerical or Likert scale ratings, or comparative judgements. However, several factors limit the accuracy and efficiency of such approaches. Constrained-decoding for format control can introduce artifacts, and unintentional biases can be introduced by prompts. Overly verbose reasoning can obscure or misalign core judgements. More fundamentally, black-box approaches can lead to untrustworthy or factually incorrect generations, frequently stemming from biases learned during pretraining (Weidinger et al., 2021; Park et al., 2023; Evans et al., 2021; Hendrycks et al., 2021).

Previous work, such as Liu et al. (2024b), demonstrates that reformulating direct-scoring tasks as ranking problems based on pairwise preferences results in better alignment with human expert labellers. To extract even more accurate judgements than these generation-based approaches, we propose using classifying probes through pairwise comparisons. Empirical work suggests models' *latent knowledge*, independent of the biases considered above, can be identified through the use of trained classifier heads on the activations of a given model. Such probes can be trained in a supervised (Alain and Bengio, 2017; Marks and Tegmark, 2024) or unsupervised (Burns et al., 2023) fashion, and importantly, can be trained using *contrast pairs*. This involves comparing the hidden state of a model when generating different possible answers, and observing salient *contrastive features* in the change of hidden state, while controlling for irrelevant features (Burns et al., 2023; Rimsky et al., 2024). This leads both to better predictive performance and gains in efficiency.

We present the first thorough investigation of the performance of supervised and unsupervised probes for LLM-as-a-Judge tasks of pairwise pref-

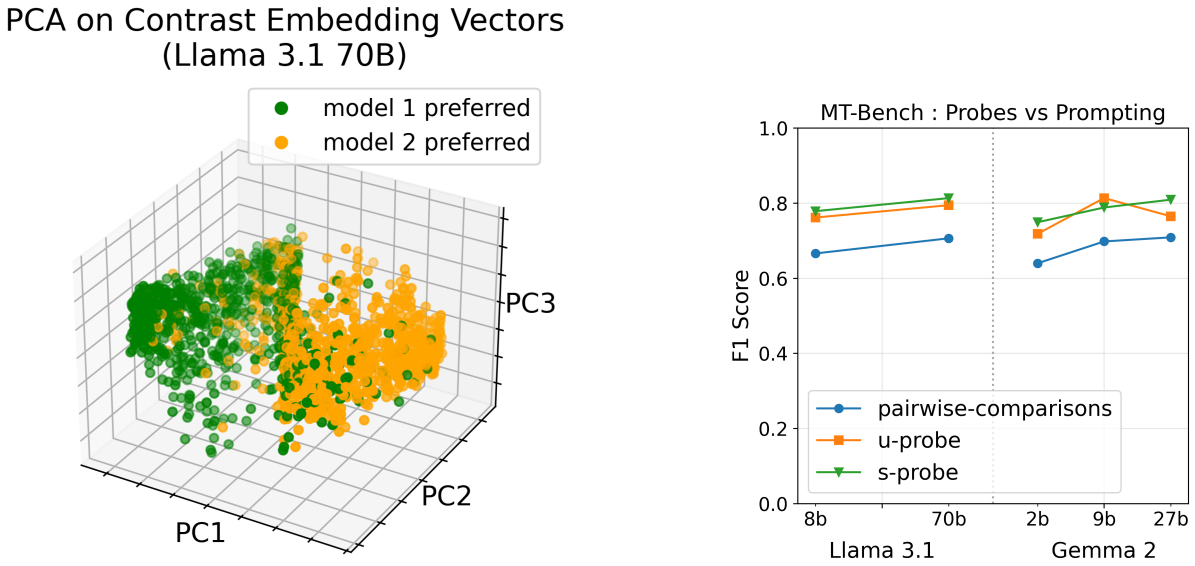

Figure 1: Our method exploits the empirical result that LLMs' internal features of "belief" or "judgement" are correlated with linear directions in their embedding spaces. For Llama 3.1 70B evaluated on the MT-Bench dataset, we find **the first principal component of the contrast pair differences of embedding vectors roughly classifies which model in a given example was preferred by a panel of human raters** *(left)*. **Supervised or unsupervised classifying probes built on these embedding vectors are more aligned with human raters than prompting methods alone** *(right)*, and this result holds across different model families (Gemma 2, Llama 3.1) at different sizes (from 2B to 70B parameters).

erences, comparing these methods to generation-based evaluation, with and without supervised fine-tuning (SFT). To summarise our main contributions:

- We introduce a way to extract human-aligned judgement from LLMs, by leveraging linear classifying probes and contrast pairs, in both a supervised and unsupervised setup.
- Through extensive experiments, we show classifying probes consistently outperform generation-based evaluation.
- We also show supervised probes considerably improve on the cost:performance ratio of SFT in realistic scenarios.
- We demonstrate these probes correlate with interpretable features of the underlying language model, generalising well to different domains and remaining more robust to distributional shifts than prompting.

Our results are consistent across four widely-used open-weights model families, at sizes ranging from 0.5B to 123B parameters, and six different datasets. In light of our results, we encourage practitioners to make use of classifying probes for similar tasks for a more cost-efficient, robust, and performant solution.

2 Background and Related Work

2.1 LLM-as-a-Judge

LLMs are increasingly employed as automatic, reference-free evaluators for assessing natural language tasks (Zhong et al., 2022; Chen et al., 2023; Wang et al., 2023; Tan et al., 2024). Their applications span a wide range of domains, such as summarisation (Shen et al., 2023), instruction following (Zeng et al., 2023), legal analysis (Deroy et al., 2024), reasoning (Stephan et al., 2024), and recommendation systems (Hou et al., 2024). Despite their growing adoption, LLM-as-a-judge faces several challenges, including misalignment with human judgments (Chiang and Lee, 2023), biases in various forms (Zheng et al., 2023b; Zhou et al., 2024), inconsistencies in decision-making (Liu et al., 2024a), and limitations in reasoning capabilities (Stephan et al., 2024).

To address these issues, researchers have proposed several methods to enhance the reliability and accuracy of LLM-based judgments. G-Eval (Liu et al., 2023) refines scoring granularity using a logit-weighted average of score tokens. Pairwise comparison techniques have been introduced to improve alignment with human preferences, as demonstrated by Liu et al. (2024b) and Liusie et al.

(2024). Other approaches advocate for generating Chain-of-Thought rationales prior to evaluation (Saha et al., 2025; Wang et al., 2024; Ankner et al., 2024) or employing multi-agent debate frameworks (Chan et al., 2024) or evaluator panels (Verga et al., 2024) to enhance assessment robustness.

2.2 Representation Probing

Probing methods assess the extent to which language model representations encode specific knowledge. Typically, a probe is a supervised classifier (Conneau et al., 2018; Hupkes et al., 2018) trained to extract information from these representations (Christiano et al., 2021; Belrose et al., 2023). Linear probing (Alain and Bengio, 2018), which employs a linear classifier, is particularly valued for its efficiency and interpretability. Unsupervised variants have also been explored (Burns et al., 2023; Laurito et al., 2024).

Probing has been widely applied to interpret model representations across various domains, including word embeddings (Levy and Goldberg, 2014), sentiment (Maas et al., 2011), factual knowledge (Marks and Tegmark, 2024), spatial and temporal understanding (Gurnee and Tegmark, 2024), and world models (Li et al., 2023). It has also been used to detect behavioural patterns such as outliers (Mallen et al., 2024), inactive modules (MacDiarmid et al., 2024), and unfaithful generation (Azaria and Mitchell, 2023; Campbell et al., 2023).

In this work, we employ linear probing to extract evaluation judgments from an LLM-as-a-judge setup. Compared to inference-based or logits-based judgments, we show that linear probing improves both accuracy and efficiency.

3 Methodology

In order to identify an LLM's true "judgement" via classifying probes, we seek to identify binary features of belief or knowledge: a given text may or may not be consistent with the knowledge the LLM has learned during training, and we wish to identify a linear direction in activation space correlated with this property.

To identify such a direction, we make use of **contrast pairs** (Burns et al., 2023). We begin with a diverse set of binary statements or questions $S = \{s_i\}_{i=1}^N$. The contrast pairs are a dataset of prompts $X = \{(x_i^+, x_i^-)\}_{i=1}^N$ constructed by appending contrasting tokens to each s_i. Suppose for example that $s_i =$ "The capital of France is Paris." A contrast pair for factual accuracy on s_i would have $x_i^+ =$ "The capital of France is Paris. This statement is true" and $x_i^- =$ "The capital of France is Paris. This statement is false".

Both prompts are then used as inputs to an LLM, and the embedding vectors $\phi(x_i^+)$ and $\phi(x_i^-)$ of the differing contrasting tokens are harvested at a layer l.

We assume both $\phi(x_i^+)$ and $\phi(x_i^-)$ can be decomposed into several features, most of which are shared (since both are derived from the statement s_i). We also assume we can approximate both as a linear combination of said features:

$$\phi(x_i^+) = \sum_{i=1}^{n} \mathcal{F}_i^{shared} + \sum_{j=1}^{m} \mathcal{F}_i^+ + \epsilon^+,$$

$$\phi(x_i^-) = \sum_{i=1}^{n} \mathcal{F}_i^{shared} + \sum_{j=1}^{k} \mathcal{F}_i^- + \epsilon^-,$$

with all $\mathcal{F}^{shared}$ common to both embeddings, $\mathcal{F}^{+/-}$ unique to each element of the contrast pair and remaining information $\epsilon^{+/-}$.

Consider the contrast pair difference $\phi(x_i^+) - \phi(x_i^-)$, removing the effect of all $\mathcal{F}^{shared}$ and leaving only **contrastive features**. Two immediately obvious contrastive features are:

- $\Delta_{syntax} := \mathcal{F}^{True} - \mathcal{F}^{False}$, the syntactical difference in the prompts x^+ and x^-.
- $\Delta_{knowledge} := \mathcal{F}^{\top} - \mathcal{F}^{\perp}$, the logical difference between both prompts: one is consistent with the model's internal knowledge while the other is not. This can be thought of as the model's "belief" in a sense.

Given a dataset of contrast pair differences $D = \{\phi(x_i^+) - \phi(x_i^-)\}_{i=1}^N$, a centering step can be performed to remove Δ_{syntax} before taking this difference:

$$\tilde{\phi}(x_i^+) := \phi(x_i^+) - \mu^+,$$

$$\tilde{\phi}(x_i^-) := \phi(x_i^-) - \mu^-,$$

where μ^+ and μ^- are the mean embedding vectors of $\{\phi(x_i^+)\}$ and $\{\phi(x_i^-)\}$ respectively. We claim $\Delta_{knowledge}$ will, in most cases, be the *most salient* contrastive feature of the new dataset $\tilde{D} = \{\tilde{\phi}(x_i^+) - \tilde{\phi}(x_i^-)\}_{i=1}^N$.

Given ground truth labels for each pairwise comparison, we can model the probability with a classifier:

$$\mathbb{P}(x^+ \text{ true}) = \sigma(\mathbf{w}^T(\tilde{\phi}(x_i^+) - \tilde{\phi}(x_i^-))).$$

The supervised probes we train in Section 4 are described by the above classifier, the parameters of which we fit using logistic regression.

Additionally, were the above claim of salience of $\Delta_{knowledge}$ to be true, it would be identifiable as the top principal component of our dataset, thereby allowing us to obtain a probing classifier without the need for ground-truth labels. Indeed, this approach is commonly used as an unsupervised probing technique for similar tasks (Burns et al., 2023; Laurito et al., 2024), and we explore its use for LLM-as-a-Judge pairwise comparisons here too.

4 Experimental Setup

4.1 Datasets

We investigate the performance of classifying probes for LLM-as-a-Judge tasks through the use of several datasets spanning different problem domains.

LLM Chat Preferences MT-Bench (Zheng et al., 2023a) is a multi-turn question set of pairwise comparisons of chatbot LLM interactions. A subset of these comparisons have been performed by several human labellers, and we use these as ground-truth to measure performance against.

Text Quality The NEWSROOM (Grusky et al., 2018), SummEval (Fabbri et al., 2020), and HANNA (Chhun et al., 2024) datasets all concern the evaluation of text in terms of high-level concepts. News articles with summaries (former two) and story prompts with short stories (last) are evaluated by human labellers on several high-level features such as "coherence" or "surprise". Note these are directly scored on a Likert scale; when necessary we convert this task to one of pairwise comparisons, following Liu et al. (2024b).

Common Sense Reasoning The ROCStories dataset (Mostafazadeh et al., 2016a) consists of short story prompts provided with two potential endings. One ending is always more consistent with the story prompt, allowing for a pairwise comparison task. We additionally make use of the MCTACO (Zhou et al., 2019) and CaTeRS (Mostafazadeh et al., 2016b) datasets which similarly test common sense reasoning in the context of causal/temporal understanding.

4.1.1 Prompts

Full prompt templates for all datasets can be found in Appendix E or in our code repository to be released. However, the general prompt template used in all pairwise comparison experiments is shown below:

```
Consider the following two <items>:
Choice 1: <item 1>
Choice 2: <item 2>
Which is more <task>?
Answers must be a single choice.
```

When harvesting contrast pairs, we prime the model with the following message:

```
Between Choice 1 and Choice 2, the more
<task> <item> is Choice <contrast_token>
```

4.2 Models

Our results robustly generalise between different LLM model families. We demonstrate this in Section 5, where we conduct scaling analyses only within model families, as idiosyncratic differences between them lead to different patterns of performance and scaling. Specifically, we consider:

- the two smaller (8B, 70B) Llama 3.1 models (Meta AI, 2024).
- the Gemma 2 (2B, 9B, 27B) family of models (Google Gemma Team, 2024).
- the Qwen 2.5 (0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B) family of models (Qwen, 2025).
- Mistral Nemo (12B), Small (22B), and Large (123B) (Mistral AI, 2024b,c,a).

All our experiments are conducted through question-answering, and due to this **all** models we use have undergone a post-training pipeline of (usually) instruction-tuning and some form of preference learning such as reinforcement learning from human feedback (Christiano et al., 2017).

4.3 Baselines

For the text quality datasets mentioned in Section 4.1 we report baseline results of generation-based prompting a given model to evaluate text on the original Likert scale e.g., on a scale of 1 to 5, referring to this as *direct-scoring*. We additionally report a recent improvement to this approach, G-Eval (Liu et al., 2023), which re-weights predictions using the model's own predicted probabilities for each possible answer choice.

When re-framing the above tasks as pairwise comparisons, and with all other datasets, we report prompting performance for comparisons. To address positional bias, we marginalise over position and take an average of the model's predicted

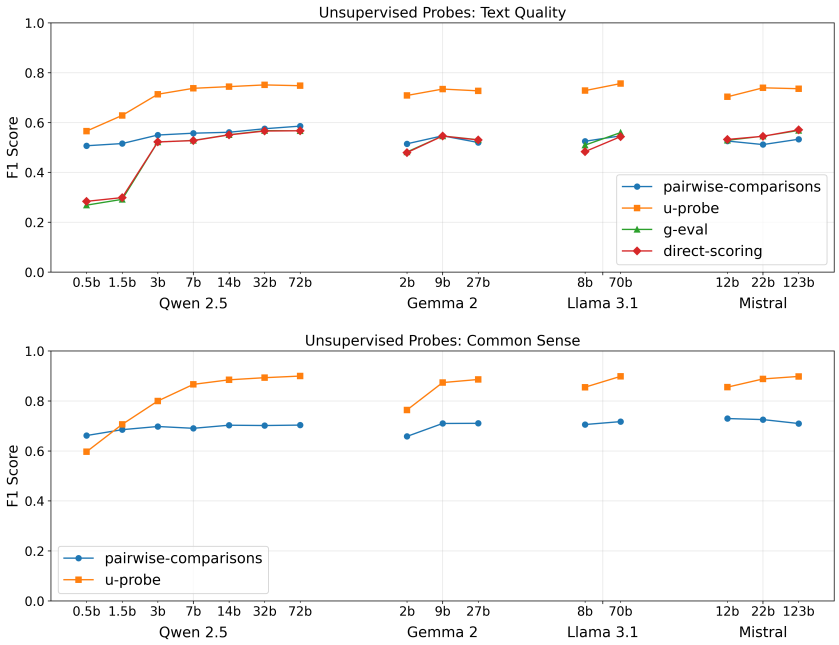

Figure 2: **Unsupervised probes, in all but one test case, outperform generation-based methods like direct-scoring and pairwise comparisons.** Interestingly, within a given model family, unsupervised probing performance with a small model almost always outperforms prompting performance with much larger models. This highlights two related key findings: (1) the use of relatively large LLMs for LLM-as-a-Judge tasks may be unnecessarily computationally wasteful and (2) there may be significant capability "left on the table" with smaller LLMs for such tasks.

probabilities. Note: this assumes a consistent positional bias, and requires us to run each question through the model twice (with the two possible answer choices swapped).

4.4 Training Setup

Generation-based prediction is performed by examining model predicted probabilities for possible answer choices e.g., for pairwise comparisons, we compare the probabilities for the tokens "1" and "2". For activation harvesting, unless otherwise stated, we take the embedding vector of the final token (that is, the contrasting token) of a given prompt, after the last decoder block and before the final normalisation layer[1].

Both supervised and unsupervised probes are fit and tested on random distinct halves of a given dataset.

5 Results

We first use the MT-Bench dataset to assess the ability of different LLMs to compare two model-generated answers to a user-question in a chatbot interaction: a common task in LLM post-training. A panel of human judges reached 80% agreement on this dataset (Zheng et al., 2023a), and we obtain ground-truth labels by taking the majority-vote of this panel. Both supervised and unsupervised probes perform similarly at aligning with this ground-truth, achieving F1 scores of roughly 0.8, as can be seen in Figure 1. Importantly, we find classifying probes outperform prompting methods, while maintaining the same inference cost of two forward passes per example. This motivates our deeper investigation into the potential of classifying probes for similar tasks, which we present now.

5.1 Experiment 1: Unsupervised Probes

We analyse the performance of the PCA-based unsupervised probing method described in Section 3 through the three **text quality** datasets NEWSROOM, SummEval, and Hanna, and the three **common sense reasoning** datasets ROCStories, MCTACO, and CaTeRS. A baseline for our probes should also be unsupervised; we compare against zero-shot prompting on all datasets, calibrating model predictions by running each input example twice, swapping the order of examples in a given pairwise comparison, allowing us to marginalise over answer position to account for order effects.

[1]The choice of layer is further investigated in Appendix A.

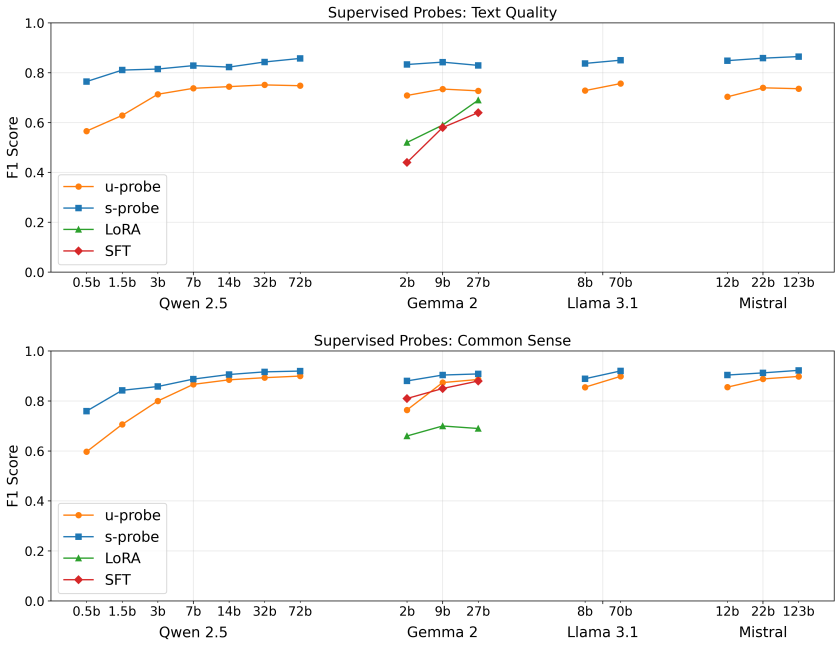

Figure 3: **Supervised probes, in all cases, allow for a further improvement in alignment with human raters over unsupervised probes.** We also test parameter-efficient and full finetuning of models in the Gemma 2 family, finding that supervised probes still outperform finetuned generation-based evaluators.

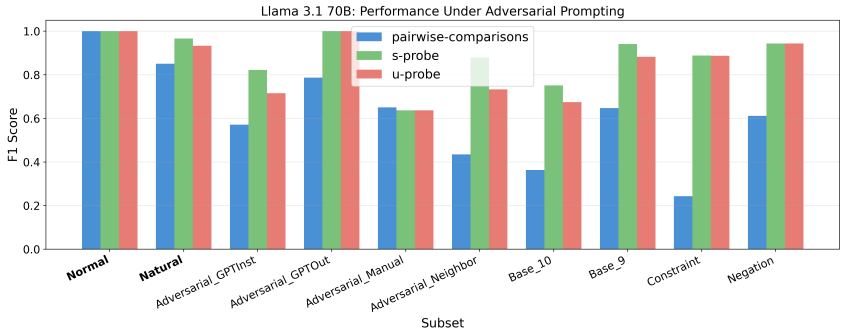

Figure 4: Performance of classifying probes and generation-based prompting for Llama 3.1 70B on the LLMBar dataset. All three methods suffer under adversarial prompting (non-bold subsets), however, **both probing approaches remain significantly more robust than prompting.**

In the case of the text quality datasets we can also report direct scoring on the original Likert scale of 1-5, and a G-Eval-based (Liu et al., 2023) correction of this approach.

Unsupervised Probes Outperform Calibrated Prompting Methods We find for all six datasets and four model families, aside from a single test case (Qwen 2.5 0.5B), the use of unsupervised probes allows for significantly higher alignment with human judgement (Figure 2). We see this as evidence of a capability-gap between models' abilities measured through the flexibility and capacity of their latent spaces and their abilities measured through standard prompting approaches. Such a gap may narrow over time with the release of newer models with higher instruction-following capabilities, but it remains sizable for now. We therefore advocate the use of unsupervised probes for pairwise comparison tasks in which labels are sparse/absent over prompting methods alone.

These results, further broken down into each of the six constituent datasets aggregated in Figure 2, can be found in Appendix F.

5.2 Experiment 2: Supervised Probes

For many LLM-as-a-Judge tasks, it may be feasible to obtain a (small) number of labelled examples

to guide the decision-making process of an LLM evaluator. In such cases, a supervised probe can be trained by replacing the PCA step above with a standard supervised classifier, as described in Section 3. For the same six datasets evaluated above, we train supervised probes on 5000 examples and examine their performance against corresponding unsupervised probes on the remaining held-out examples.

Supervised Probes Outperform Unsupervised Probes And Can Outperform Finetuning We find, as shown in Figure 3, that supervised probes often allow for a *further* increase in alignment with human raters. For particularly sensitive tasks in which supervised approaches are feasible, practitioners may opt to finetune a given model to improve its performance. We also find supervised probes are a competitive alternative to such an approach, as shown for the Gemma 2 model family in Figure 3. For both the text quality and common sense reasoning tasks, supervised probes outperform LoRA (Hu et al., 2022) and even full finetuning with the same number of training examples, at all model sizes [2].

The results in Figure 3 may shed some light on the difficulty of the text quality and common sense reasoning tasks set up in our experiments. Note for the former, finetuned models actually perform relatively poorly, with a large capability gap against both unsupervised and supervised probes. In the common sense reasoning task, finetuning is much more competitive. We hypothesize this is due to the subjective vs objective nature of the two tasks. The evaluation of text on abstract features such as "coherence" and "empathy" (as is carried out in the NEWSROOM, SummEval, and HANNA datasets) is likely highly subjective, while common sense reasoning can be considered a much more objective task. This makes the latter much easier to learn during pretraining, and further improve on during finetuning. This is further reflected in the larger improvement in supervised over unsupervised probes for the text quality task: human-generated labels allow the probe-fitting process to efficiently align with raters. Conversely, finetuning approaches likely require many more labels to converge to this same distribution, with orders of magnitude more parameters requiring tuning over the logistic regression classifiers we train.

[2]Details on the finetuning process performed are provided in Appendix D.

This suggests a key advantage of probing approaches constructed through contrast pairs: the salience of the desired knowledge or belief is increased, facilitating easier learning of the task and reductions in computational cost compared to finetuning. We expect, in the limiting case of labelled data, finetuning approaches will overtake probe performance due to their higher flexibility. However, for many realistic applications, labelled data can be unreasonably expensive to obtain.

These results, broken down into each of the six constituent datasets aggregated in Figure 3, can be found in Appendix F.

5.3 Experiment 3: Probe Generalisation

In addition to offering advantages in both computational cost and performance for LLM-as-a-Judge tasks, classifying probes yield key interpretability insights into LLMs in general. We find evidence they correlate with *general* features of belief or judgement used by a given model.

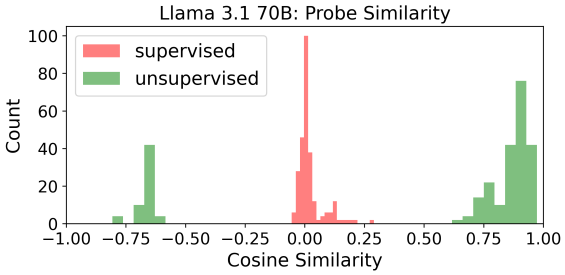

Figure 5: Taking the example of Llama 3.1 70B, we find most supervised probes are dissimilar while most unsupervised probes are similar (up to sign), regardless of the varying tasks in each of the six datasets considered.

This evidence, summarised in Table 1, comes from an experiment into the generalisation of probes under significant distributional changes. Specifically, we train both supervised and unsupervised probes using contrast pair differences of activations from Llama 3.1 70B on each of the six datasets examined in the above experiments, and test each on the remaining five other datasets.

Classifying Probes Identify Generalising Features Of Belief Or Judgement Generally, these probes achieve high F1 scores, even when trained and tested on very different tasks. Unsupervised probes in particular generalise extremely well in several cases, achieving F1 scores at or above 0.95. We hypothesize the contrast pair setup allows probes to focus on *task-independent* features of

F1-Score Supervised/Unsupervised		Probe					
		NEWSROOM	SummEval	HANNA	ROCStories	MCTACO	CaTeRS
Evaluation	NEWSROOM	-	0.63/0.77	0.65/0.77	0.78/0.77	0.78/0.77	0.64/0.77
	SummEval	0.66/0.76	-	0.63/0.76	0.77/0.76	0.77/0.76	0.58/0.76
	HANNA	0.67/0.71	0.62/0.71	-	0.71/0.71	0.71/0.71	0.59/0.70
	ROCStories	0.87/0.99	0.78/0.99	0.79/0.98	-	0.99/0.99	0.71/0.99
	MCTACO	0.79/0.95	0.67/0.95	0.70/0.95	0.96/0.95	-	0.74/0.95
	CaTeRS	0.75/0.78	0.62/0.78	0.67/0.78	0.76/0.78	0.77/0.78	-

Table 1: **Generalisation of classifying probes for Llama 3.1 70B.** We train both supervised *(left)* and unsupervised *(right)* probes on examples from a given dataset *(columns)*, testing them on all five other datasets *(rows)* through F1-score. The higher scoring probe of a given supervised/unsupervised pair is coloured. We find both sets of probes generalise relatively well, and unsupervised probes in particular generalise very well on several occasions.

judgement, relying on models' already vast knowledge and human alignment due to pretraining to infer the correct choice. Supervised probes can leverage information about a specific task, but this ultimately pushes the probe direction away from these task-independent features, leading to slightly worse generalisation than their unsupervised variants.

The better generalisation of unsupervised probes is supported by Figure 5, in which the cosine similarity between all probe directions across all datasets is plotted. Up to a sign flip, unsupervised probes are highly similar, with most having magnitude similarity above 0.7. Meanwhile, the distribution for supervised probes is narrowly centred around zero. It is possible that supervised probe similarity could be increased by training on more/diverse data, but it is striking that unsupervised probes trained on relatively different domains identify similar features. It is unclear however whether these features are causally relevant during the forward pass, to represent belief and judgement, but we investigate this further in Appendix B.

5.4 Experiment 4: Ablation Study

As a final test of classifying probes, we perform an ablation study of performance under different types of adversarial prompting strategies. To do so, we make use of the LLMBar dataset (Zeng et al., 2024) for evaluating instruction-following capabilities. This dataset is split into several subsets, all of which, other than the *Normal* and *Natural* subsets, are specifically designed to induce incorrect answers from LLM evaluators.

Classifying Probes Are More Robust to Domain Shifts Than Prompting We train probes on the *Normal* and *Natural* subsets only, testing them on all other subsets, and comparing with generation-based prompting as before. Figure 4 shows, for Llama 3.1 70b, how all methods suffer a performance drop under adversarial prompts. However, we note for all but one subset, this drop is significantly less severe for probing approaches over prompting; note in particular the results on the *Constraint* subset for example. This finding holds across different model sizes and families - our replications of this experiment on other models can be found in Appendix C. This complements our results on probe generalisation: the relative robustness of classifying probes likely aids their ability to generalise to different domains.

6 Conclusion

We explore the use of linear classifying probes, both supervised and unsupervised, to perform pairwise comparisons in several standard LLM-as-a-Judge tasks. Our approach of using contrast pair differences to increase the salience of relevant "belief" or "judgement" features proves to be greatly effective; unsupervised probes consistently outperform calibrated generation-based evaluators across several open-weights LLM families and model sizes, without a significant increase in computational cost. In realistic scenarios with limited but available ground-truth labels, we also find supervised probes outperform unsupervised methods and can even outperform finetuning of the same model. These probes generalise well to different domains, and are more robust to (adversarial) distributional shifts than prompting approaches. Our experiments constitute the first comprehensive assessment of both supervised and unsupervised probes for LLM-as-a-Judge tasks against generation-based approaches, with and without finetuning. They suggest for practical applications, classifying probes are a cost-efficient, robust, and powerful solution.

Limitations

Our experiments in Section 5.2 find supervised probes outperform finetuned (both LoRA (Hu et al., 2022) and full) generation-based evaluators given the same training data. It could be interesting to investigate how and when probe performance saturates, and relatedly whether finetuning approaches outperform probes in the limit of data availability. While this is outside the scope of our study, future work establishing the threshold at which this may (or may not) take place would better inform developers of best practices.

Additionally, we focus the scope of LLM-as-a-Judge tasks covered in this work to those of pairwise preferences, as this setup lends itself well to the use of binary classifying probes. We would be excited to see future work exploring the use of latent knowledge in direct-scoring tasks, where texts are rated on a numerical or Likert scale. This could be achieved through one-vs-rest or multi-class probes for example.

Finally, there remain additional challenges to overcome with probing methods in general. Red-teaming studies and analyses (Farquhar et al., 2023; Laurito et al., 2024) find that prompts which can induce a language model into simulating a different *quality* of knowledge e.g., *"You are a smart professor..."*, can significantly affect probe performance. Addressing this challenge proves to be particularly difficult for the research community, as it requires a much better understanding of knowledge representation within LLMs. For now, this presents a fundamental limitation of probing and other similar white-box approaches.

Responsible NLP Statement

We follow strict compliance with all dataset and model licenses relevant in this work. AI assistants were used in the process of writing experimental code only.

Acknowledgments

[anonymized]

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan Daniel Chang, and Prithviraj Ammanabrolu. 2024. Critique-out-loud reward models. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

James Campbell, Phillip Guo, and Richard Ren. 2023. Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching. In *Socially Responsible Language Modelling Research*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? Prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Paul Christiano, Ajeya Cotra, and Mark Xu. 2021. Eliciting latent knowledge: How to tell if your eyes deceive you. *Google Docs, December*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, pages 1–44.

Owain Evans, Owen Cotton-Barratt, Lars Finnveden, Adam Bales, Amanda Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. 2023. Challenges with unsupervised llm knowledge discovery. *Preprint*, arXiv:2312.10029.

Google Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Walter Laurito, Sharan Maiya, Grégoire Dhimoïla, Owen Ho Wan Yeung, and Kaarel Hänni. 2024. Cluster-norm for unsupervised probing of knowledge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14083–14112, Miami, Florida, USA. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. 2024a. Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *arXiv preprint arXiv:2410.02205*.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. 2024. Simple probes can catch sleeper agents.

Alex Troy Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. 2024. Eliciting latent knowledge from "quirky" language models. In *First Conference on Language Modeling*.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Meta AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mistral AI. 2024a. Mistral large. https://mistral.ai/en/news/mistral-large.

Mistral AI. 2024b. Mistral nemo. https://mistral.ai/en/news/mistral-nemo.

Mistral AI. 2024c. Mistral small 3. https://mistral.ai/en/news/mistral-small-3.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Peter S Park, Samuel Goldstein, Annette O'Gara, Mingjie Chen, and Dan Hendrycks. 2023. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*.

Qwen. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. 2024. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. *arXiv preprint arXiv:2409.04168*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, et al. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. *arXiv preprint arXiv:2401.15042*.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-taught evaluators. *Preprint*, arXiv:2408.02666.

Laura Weidinger, John Mellor, Moritz Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Michael Cheng,

Matthew Glaese, Borja Balle, Atoosa Kasirzadeh, and Zachary Kenton. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

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

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Han Zhou, Xingchen Wan, Yinhong Liu, Nigel Collier, Ivan Vulić, and Anna Korhonen. 2024. Fairer preferences elicit improved human-aligned large language model judgments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1241–1252, Miami, Florida, USA. Association for Computational Linguistics.

A On The Target Layer For Activation Harvesting

All classifying probes examined in Section 5 are trained following the same process. One key step, as explained in Section 3, involves the harvesting of activations during the forward pass. In the contrast pair setup, we do so on the token position of contrasting tokens at the last layer of a given model. Here, we briefly explore how necessary this choice is.

We train supervised and unsupervised probes on models from the Gemma 2 and Llama 3.1 families on the MT-Bench dataset by harvesting activations on contrasting tokens at *all* layers of a given model. The performance of the downstream trained probes are compared in Figure 6 (supervised probes) and Figure 7 (unsupervised probes).

A comparison of the two figures is striking, and reveals some of the key differences between supervised and unsupervised linear probes. For both, average performance is poorer when probes are trained on the first half of a given model than the second. For supervised probes, the increase in performance is smooth: by around 40-50% of the way through the forward pass, they are able to learn as best they can for the given task.

In contrast, we see a discontinuity in unsupervised probe performance. This discontinuity appears after different numbers of layers depending on the model, but we note the larger the model the earlier it appears (for a given model family). This discontinuity sees performance jump from an F1 score of roughly 0.5 (a balance between precision and recall, moderately better than random classification) to a maximum of roughly 0.8.

Taking our results in Section 5.3 at face value, we hypothesize the main reason for this difference is the salience of the desired feature. Supervised probes are in a sense reflecting a quality of the latent space itself, and how easy/difficult it may be to identify any given feature within this space. Unsupervised probes, by design, rely on the assumption that the desired feature is the *most salient* of the contrast pair differences, rather than the existence of the feature at all.

The results in Figure 7 suggest that in larger models this quality of salience is realised earlier in the forward pass, perhaps due to higher representational capacity.

Nonetheless, our decision to harvest activations at the last layer of a given model appears justified, as performance in both Figure 6 and Figure 7 remains at its best through the last layer. For practitioners, this is particularly convenient as extraction of the last hidden state of a given model is more easily facilitated in common open-LLM frameworks than earlier layers.

B Causal Analysis Of Probe Directions

Within the context of concept-based interpretability of LLMs, the term *feature* is ill-defined. Specifically, it is unclear what exactly constitutes a "true" feature of a given model. One possible definition is causal in nature: were the ablation of a given feature representation to result in a model unable to represent said feature, it is in some sense "true" and causally relevant during the forward pass. This idea has been used to investigate and "steer" LLM features in previous works such as Arditi et al. (2024) and Rimsky et al. (2024), and we follow a similar approach here to investigate the extent to which the features identified by our supervised and unsupervised probes are "true".

We repeat the prompting experiments performed using the MT-Bench dataset with models from the Gemma 2 and Llama 3.1 families. However, for each model, we orthogonalize the last token's embedding against either the supervised or unsupervised probe directions identified before, at all layers during the forward pass. That is, after each decoder block and the final layer normalisation, we perform the vector rejection,

$$x' = x - \frac{x \cdot p}{p \cdot p} p,$$

replacing the original hidden state vector x with x' given the probe direction p, meaning the model's computational operations are never permitted to write information along this probe direction. The difference in evaluator performances (from the baseline un-steered model) are shown in Table 2.

We see negligible change in evaluator performance regardless of probe type, and consider this evidence against the hypothesis that the features classifying probes identify are true features used by a given language model during LLM-as-a-Judge evaluation.

This may be due to the nature of high-dimensional space: it is likely there are several high-performing linear classifiers for such a task, and our probes are only ever able to identify one.

It may also be the case that features used for the *expression* of a belief are different from those used

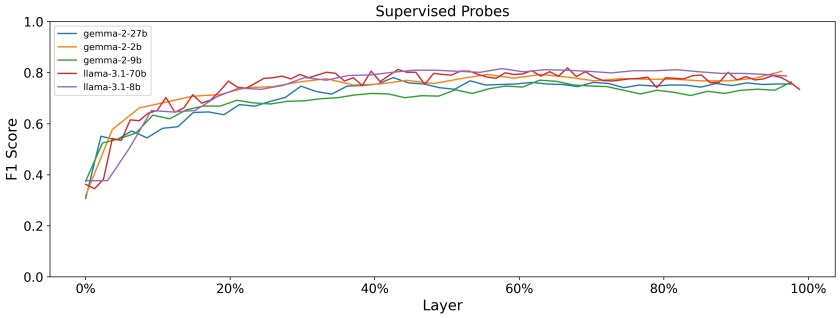

Figure 6: Performance of supervised probes trained on all layers of a given language model and evaluated on the MT-Bench dataset. We see a relatively **smooth increase in probe performance through the forward pass.**

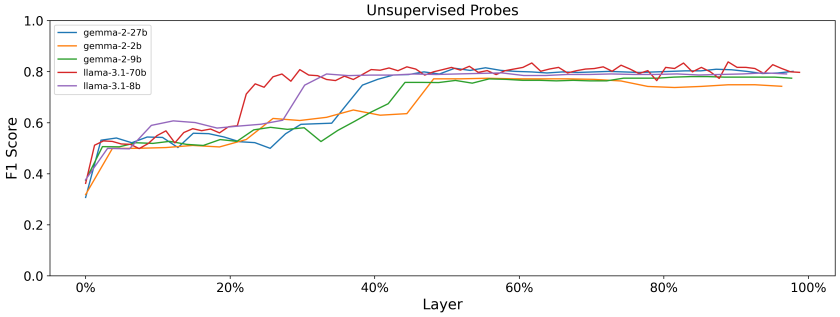

Figure 7: Performance of unsupervised probes trained on all layers of a given language model and evaluated on the MT-Bench dataset. In contrast with Figure 6, we see a **discontinuous jump in performance of unsupervised probes at differing points during the first half of the forward pass.**

to assess belief in a token *already generated*. Note in the contrast pair setup, activations are harvested at the contrasting token position, as opposed to the token before. It may well be that our probes, in particular our unsupervised probes, are identifying features related to a model's belief in its own utterances, rather than features related to evaluation itself. This raises intriguing questions regarding how realistic off-policy vs on-policy experiments with LLMs are, and we would be excited to see this explored in future work.

C Additional Results For Our Ablation Study

We repeat the experiments performed in Section 5.4 on Gemma 2 2B, 9B, 27B, and Llama 3.1 8B. Results are consistent with our tests on Llama 3.1 70B in that probes are, in general, more robust to adversarial prompting strategies than generation-based inference. Note this is particularly apparent with the smallest model tested (Gemma 2 2B).

Results are shown in Figure 8 through to Figure 11.

D Details Of Supervised Finetuning Experiments

For the experiments in Section 5.2 in which we compare supervised probe performance with finetuning for models Gemma 2 2B, 9B, and 27B. We use the OpenRLHF (Hu et al., 2024) library. For LoRA (Hu et al., 2022) finetuning we use a rank of 64 and α of 64, targeting all modules. In all cases we train on a dataset of 5000 randomly chosen samples for one epoch. Full training configs will be available in our code repository to be published.

E Full Prompts for All Datasets

E.1 Text Quality Datasets

The text quality datasets NEWSROOM (Grusky et al., 2018), SummEval (Fabbri et al., 2020), and HANNA (Chhun et al., 2024) all present the task of the evaluation of generated text on high-level, abstract features or aspects. Each original dataset includes descriptions of these features, which we use in our evaluator prompts for additional context. These descriptions are listed in Table 3.

The prompt formats for the NEWSROOM and SummEval datasets follow a very similar structure,

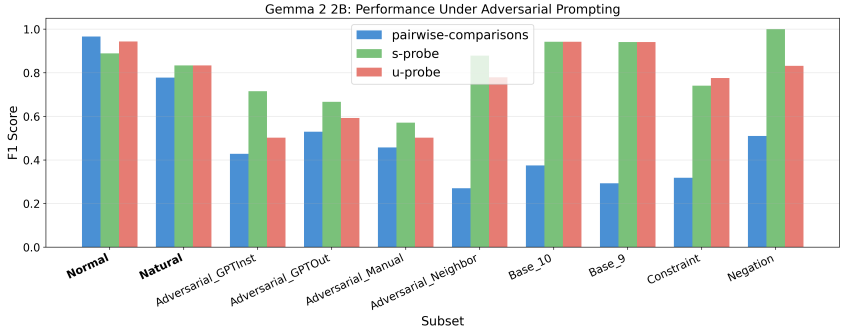

Figure 8: Performance of classifying probes and standard prompting for Gemma 2 2B on the LLMBar dataset.

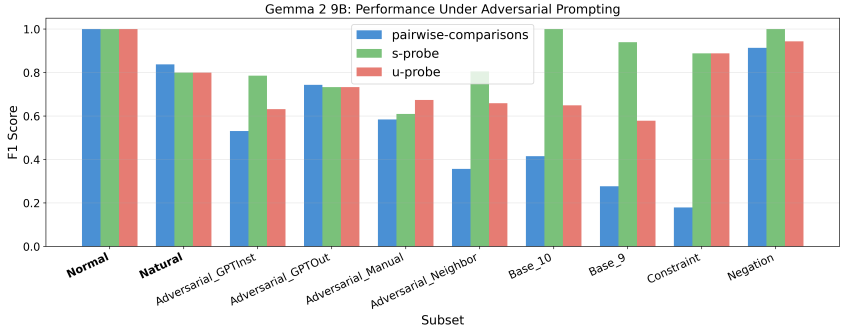

Figure 9: Performance of classifying probes and standard prompting for Gemma 2 9B on the LLMBar dataset.

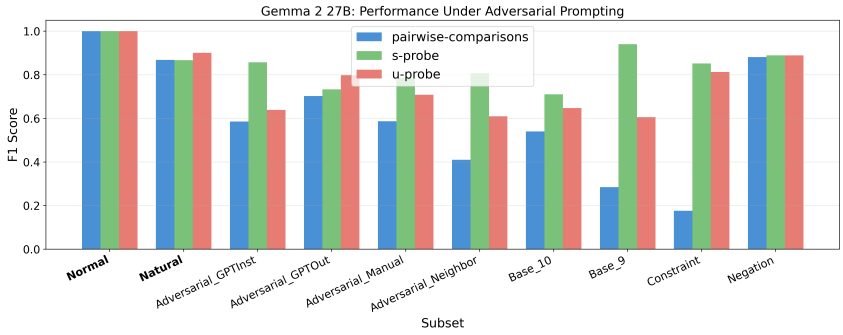

Figure 10: Performance of classifying probes and standard prompting for Gemma 2 27B on the LLMBar dataset.

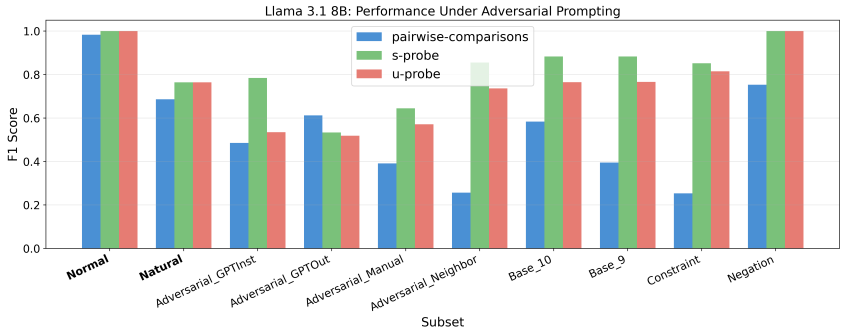

Figure 11: Performance of classifying probes and standard prompting for Llama 3.1 8B on the LLMBar dataset.

Δ F1 Score	Gemma 2 2B	Gemma 2 9B	Gemma 2 27B	Llama 3.1 8B	Llama 3.1 70B
Supervised	-0.00	0.01	0.00	0.00	0.03
Unsupervised	-0.01	0.01	0.00	0.01	0.03

Table 2: Change (Δ F1) in evaluator performance on the MT-Bench dataset following the ablation of a given probe direction during the forward pass. For all models tested, we see neglibible change in the model's capability when it is unable to write information against the probe direction, suggesting these directions are not causally relevant for evaluation.

as both assess the same exact task: a news article CONTEXT is provided with given summaries (ITEMs). We include the relevant DESCRIPTION according to the ASPECT under study, consulting Table 3. For the direct-scoring setting, the prompt template used for NEWSROOM is:

```
Consider the following article and summary:
Article: {CONTEXT}
Summary: {ITEM}
{DESCRIPTION} Rate the {ASPECT} of this
summary from 1 to 5, where 1 represents
very low {ASPECT}, and 5 represents
excellent {ASPECT}. Responses must be a
single score.
```

For SummEval, the template is changed slightly, to match the original dataset and paper:

```
Consider the following source and summary:
Source: {CONTEXT}
Summary: {ITEM}
{DESCRIPTION} Rate the {ASPECT} of this
summary from 1 to 5, where 1 represents
very low {ASPECT}, and 5 represents
excellent {ASPECT}. Responses must be a
single score.
```

For pairwise comparisons, we follow a very similar template. For NEWSROOM:

```
Consider the following article:
Article: {CONTEXT}
Below are two summaries of the above
article:
Summary 1: {ITEM1}
Summary 2: {ITEM2}
{DESCRIPTION} Which summary is more
{ASPECT}? Responses must be a single
choice.
```

And for SummEval:

```
Consider the following source:
Source: {CONTEXT}
Below are two summaries of the above
source:
Summary 1: {ITEM1}
Summary 2: {ITEM2}
{DESCRIPTION} Which summary is more
{ASPECT}? Responses must be a single
choice.
```

For the HANNA dataset, we evaluate stories generated from story-prompts. The above template is therefore adjusted slightly. For direct-scoring:

```
Consider the following prompt and story:
Prompt: {CONTEXT}
Story: {ITEM}
{DESCRIPTION} Rate the {ASPECT} of this
story from 1 to 5, where 1 represents
very low {ASPECT}, and 5 represents
excellent {ASPECT}. Responses must be a
single score.
```

And for pairwise comparisons:

```
Consider the following prompt:
Prompt: {CONTEXT}
Below are two stories inspired
by the above prompt:
Story 1: {ITEM1}
Story 2: {ITEM2}
{DESCRIPTION} Which story is more
{ASPECT}? Responses must be a single
choice.
```

E.2 Common Sense Reasoning Datasets

For the ROCStories (Mostafazadeh et al., 2016a), MCTACO citepzhou-etal-2019-going, and CateRS (Mostafazadeh et al., 2016b) datasets, the task is formatted only as one of pairwise comparisons. Additionally, in all cases the evaluator must pick the more sensible option, so all prompt templates are very similar. For ROCStories:

```
Consider the following short story:
Story: {STORY}
Below are two statements:
Statement 1: {STATEMENT1}
Statement 2: {STATEMENT2}
Considering the context of the above
```

Dataset	Aspect	Description
NEWSROOM	Informativeness	Informativeness is how well a summary of an article captures the key points of the article.
	Relevance	The details provided in a relevant summary of an article are consistent with details in the article.
	Fluency	In a fluent summary of an article the individual sentences are well-written and grammatical.
	Coherence	In a coherent summary of an article the phrases and sentences fit together and make sense collectively.
SummEval	Coherence	Coherence is the collective quality of all sentences. A coherent summary of a source should be well-structured and well-organized. It should not be a heap of related information, but should build from sentence to sentence to a coherent body of information about the source.
	Consistency	Consistency is the factual alignment between a summary and summarized source. A coherent summary contains only statements that are entailed by the source document.
	Fluency	Fluency is the quality of individual sentences. A fluent summary of a source should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
	Relevance	Relevance is the selection of important content from a source. A relevant summary should include only important information from the source document.
HANNA	Relevance	A relevant story matches its prompt.
	Coherence	A coherent story makes sense.
	Empathy	An empathetic story allows the reader to understand the character's emotions.
	Surprise	A surprising story has a surprising end.
	Engagement	An engaging story allows the reader to engage with it.
	Complexity	A complex story is elaborate.

Table 3: Text descriptions of high-level abstract features/aspects in all text quality datasets. These are provided in prompts for additional context.

```
story, which statement is more
consistent? Responses must be a single
choice.
```

With each `STORY` and `STATEMENT`s obtained from the dataset directly.

Similarly for MCTACO:

```
Consider the following passage:
Passage: {PASSAGE}
Below is a question regarding
the above passage:
Question: {QUESTION}
Choice 1: {CHOICE1}
Choice 2: {CHOICE2}
Which answer is more sensible?
Responses must be a single choice.
```

This dataset assesses common sense reasoning through a specific `QUESTION` for each `PASSAGE`.

Lastly, for CaTeRS:

```
The following list of statements
form a story, however they are
unordered:
Unordered Statements: {UNORDERED}
Below are two statements from this
list:
Statement 1: {STATEMENT1}
Statement 2: {STATEMENT2}
Determine the correct order of the
above statements - which statement
```

`appears before the other? Responses must be a single choice.`

This dataset includes lists of unordered statements, with the pairwise comparison task set up of identifying the correct ordering of two such statements, thereby assessing temporal understanding.

F Probe Performance by Dataset

We present the performance of supervised (Figure 12 to Figure 17) and unsupervised (Figure 18 to Figure 23) probes on all constituent datasets of the text quality (NEWSROOM (Grusky et al., 2018), SummEval (Fabbri et al., 2020), HANNA (Chhun et al., 2024)) and common sense reasoning (ROCStories (Mostafazadeh et al., 2016a), MCTACO (Zhou et al., 2019), CaTeRS (Mostafazadeh et al., 2016b)) tasks examined in Section 5.

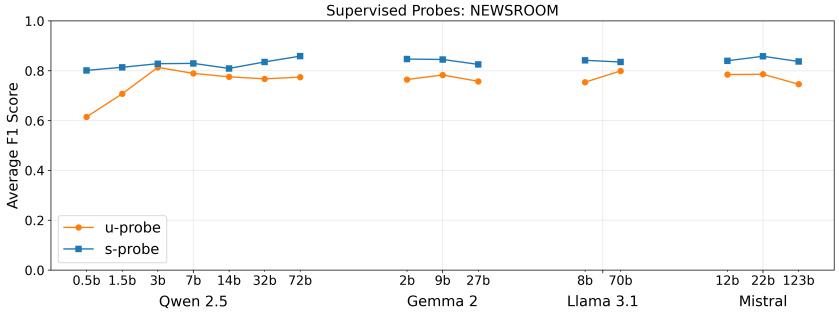

Figure 12: Supervised probe performance on the NEWSROOM dataset.

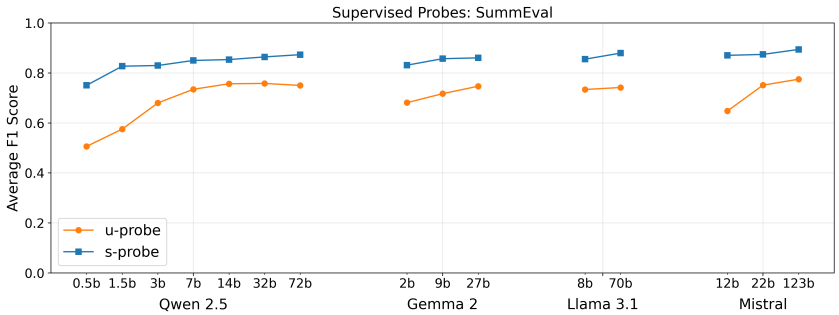

Figure 13: Supervised probe performance on the SummEval dataset.

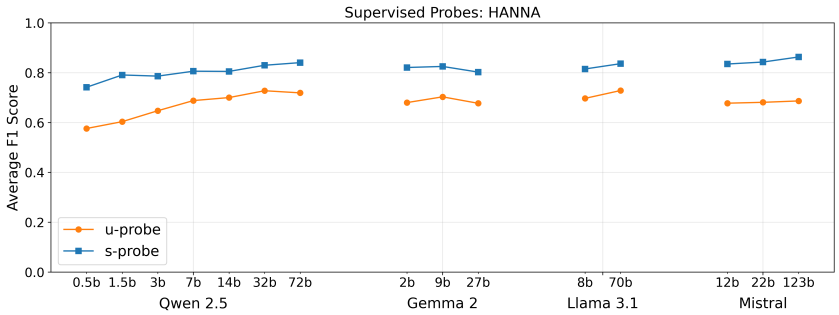

Figure 14: Supervised probe performance on the HANNA dataset.

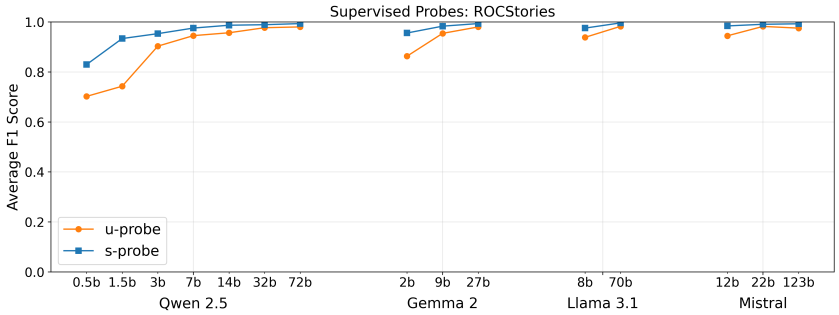

Figure 15: Supervised probe performance on the ROCStories dataset.

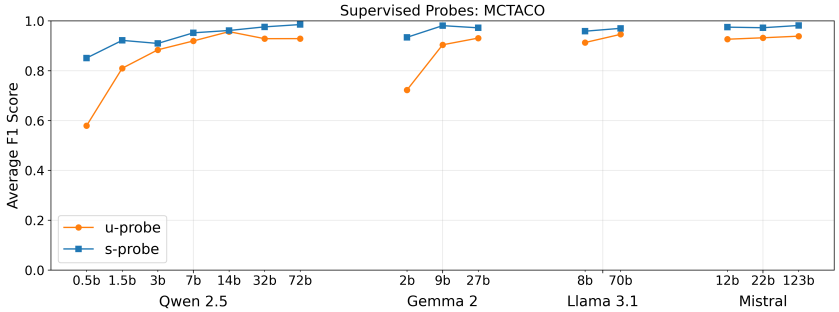

Figure 16: Supervised probe performance on the MCTACO dataset.

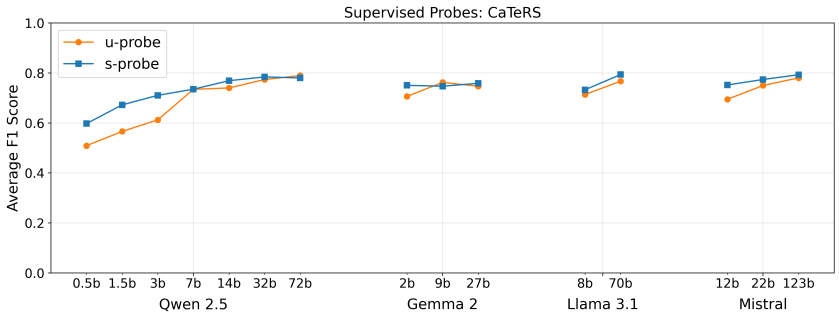

Figure 17: Supervised probe performance on the CaTeRS dataset.

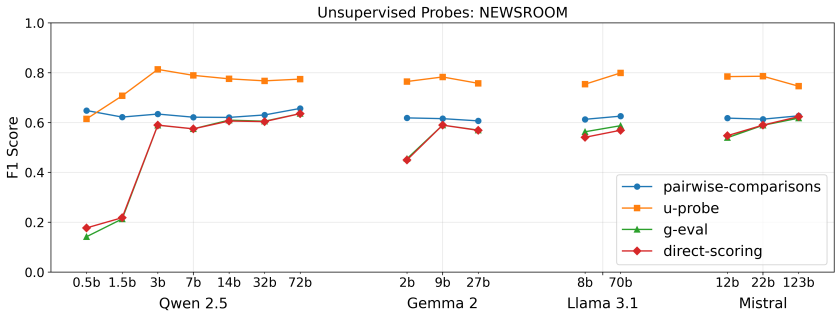

Figure 18: Unsupervised probe performance on the NEWSROOM dataset.

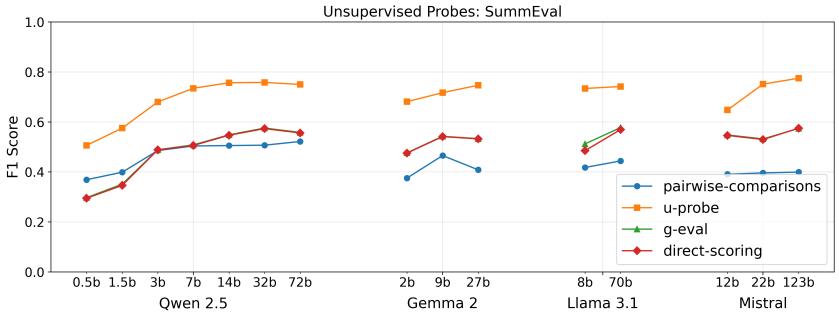

Figure 19: Unsupervised probe performance on the SummEval dataset.

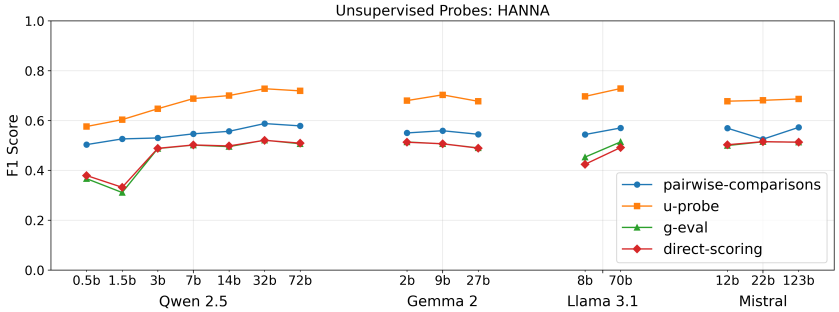

Figure 20: Unsupervised probe performance on the HANNA dataset.

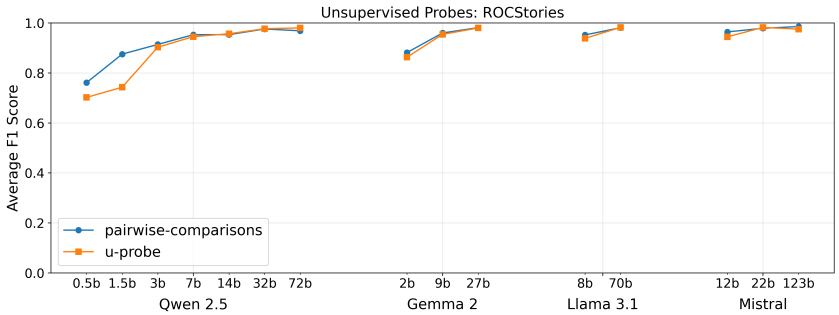

Figure 21: Unsupervised probe performance on the ROCStories dataset.

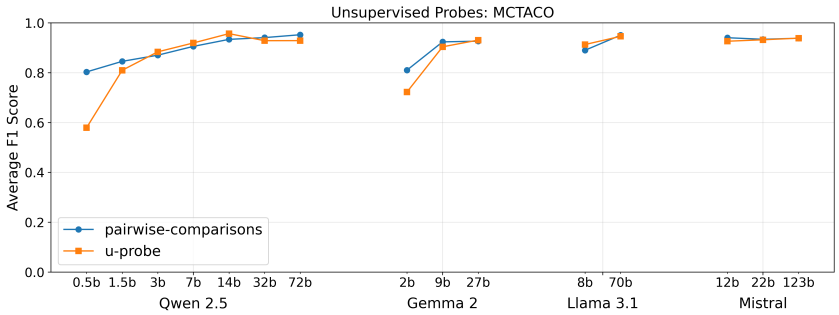

Figure 22: Unsupervised probe performance on the MCTACO dataset.

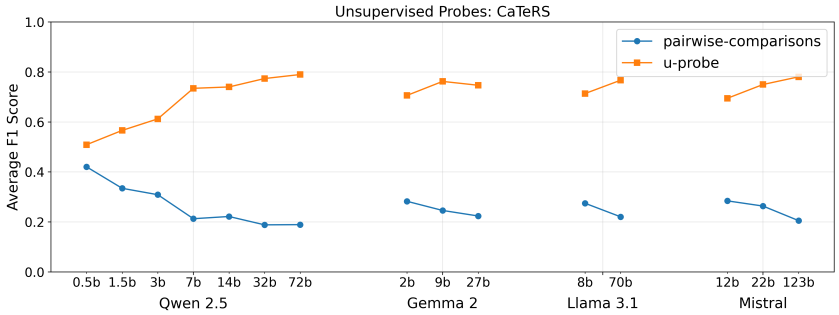

Figure 23: Unsupervised probe performance on the CaTers dataset.