Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Measuring Stability and Biases of LLM-as-a-Judge for Opinion Alignment

**Samuele D'Avenia**

samuele.davenia@unito.it

University of Turin

**Marco Madeddu**

University of Turin

**Silvia Casola**

Universitat Pompeu Fabra

**Valerio Basile**

University of Turin

**Viviana Patti**

University of Turin

**Research Article**

**Keywords:**

**Posted Date:** November 25th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11510348/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** Competing interest reported. The author Valerio Basile is a Guest Editor of the collection “Reliability and Bias in LLM-as-a-Judge Approaches” and confirms that they were not involved in the handling or decision-making process for the manuscript.

# Measuring Stability and Biases of LLM-as-a-Judge for Opinion Alignment

Samuele D'Avenia[1*†], Marco Madeddu[1*†], Silvia Casola[2], Valerio Basile[1], Viviana Patti[1]

[1]Department of Computer Science, University of Turin, Italy.
[2]Department of Engineering, Universitat Pompeu Fabra, Spain.

*Corresponding author(s). E-mail(s): samuele.davenia@unito.it; marco.madeddu@unito.it;
Contributing authors: silvia.casola@upf.edu; valerio.basile@unito.it; viviana.patti@unito.it;
†These authors contributed equally to this work.

**Abstract**

LLM-as-a-Judge is increasingly used to score how well a model-generated text aligns with a human-written reference, both as an evaluation metric and as a reward for training. This set-up is also employed for alignment with human opinions on contested topics, such as politics or moral dilemmas, to train models which are capable of reproducing the diverse viewpoints that individuals hold.
Yet, its stability and biases in this setting remain understudied: if the judge's score depends on factors other than the similarity between the two opinions, these factors become a part of what is measured and, if used as training signals, of what is rewarded. We test this through two invariances that an alignment score between two opinions should satisfy: it should not change when the judge instruction is paraphrased, and it should not change when the two opinions swap roles as candidate and reference. We evaluate 11 judges on three datasets, and relate instabilities and biases to features of the opinions being compared. We find that all judges are unstable, as swapping the order of the two opinions shifts the score by more than 30% of the rating scale in over 10% of pairs.
This instability concentrates on pairs whose opinions differ in length, and on pairs where the judge itself prefers one opinion over the other. The effects are also directional: judges score higher when the candidate is longer than the reference, and when the candidate expresses the viewpoint the judge prefers for that situation, regardless of the reference viewpoint. Used as rewards, these judges would

push models towards longer outputs and towards the judges' own viewpoints, rather than towards the human opinions they are meant to reproduce.

**Keywords:** Large Language Models, LLM-as-a-Judge, Reliability, NLG Evaluation, Meta-Evaluation, Model-Human Alignment, Opinion Bias

# 1 Introduction

The rapid progress of Large Language Models (LLMs) has made the evaluation of generated text a central challenge. Automatic metrics such as BLEU and ROUGE are cheap to compute but often fail to capture human judgements of text quality [1–3]. Human evaluation remains the de facto gold standard, but it is costly and time-consuming, which limits its use across many evaluation settings [1].

LLM-as-a-Judge has consequently emerged as a widely adopted alternative to both [4–6]. By using language models to assess nuanced and subjective properties of generated outputs, LLM-as-a-Judge aims to approximate human judgements at a fraction of the cost and time.

As LLM-as-a-Judge has been adopted across tasks, several studies have documented its limitations, including position bias, verbosity bias and other systematic biases that undermine the reliability of its judgements [4, 7, 8]. These biases could result in an evaluation that does not reflect human preferences, but is rather driven by other technical factors that are unknown and difficult to control. Moreover, recent works have also begun to use LLM-as-a-Judge scores as training signals in reinforcement learning [9–13], which raises the stakes of judge reliability. In this scenario, an unreliable or biased judge not only distorts evaluation: its biases can be passed on to the trained model, which may then inherit unwanted properties and align poorly with the gold data it is trained on, progressively shifting its distribution towards that encoded by the evaluator.

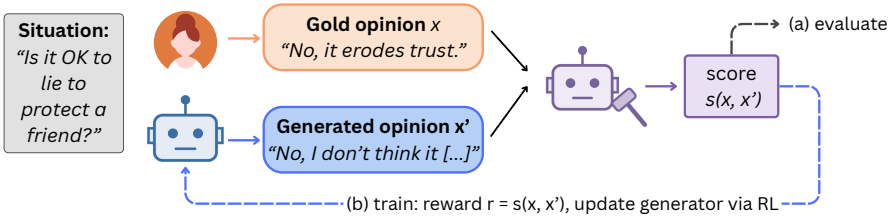



**Fig. 1**: Current usages of LLM-as-a-Judge for opinion alignment as an evaluator (a) or to obtain training reward signals (b).

LLM-as-a-Judge has also recently been adopted in settings where models are expected to reproduce human opinions on contested topics, such as politics. Several studies use LLM-as-a-Judge to assess how well model-generated opinions (candidate opinions) reflect gold-standard human opinions (reference opinions), and in some cases

use the resulting scores as training signals (Figure 1) [12–16]. In this setting, LLM-as-a-Judge may not merely measure how well a model reproduces human opinions, but actively shape which opinions are rewarded, potentially flattening genuine variation in human perspectives and suppressing minority viewpoints.[1] This is especially concerning when the resulting systems are deployed to interact with users, as LLM-generated opinions can themselves influence users' beliefs and decisions [20–22].

To our knowledge, however, no work has yet examined the **stability and biases of LLM-as-a-Judge in the opinion-alignment setting**, where the LLM-as-a-Judge is asked to assess how well a generated opinion $x'$ aligns with a human opinion $x$, and findings from existing studies on LLM-as-a-Judge bias do not transfer directly to this setting. Previous work has typically studied pairwise preferences on quality-based tasks: in contrast, judging opinion alignment is better framed as a graded score rather than a choice. Moreover, as the judges are scoring opinions, reliability might not only depend on features of the text, such as its length, but also on whether the LLM-as-a-Judge itself agrees with the viewpoint being expressed.

In this paper, we address this gap by testing LLM-as-a-Judge in the task of assigning a similarity score to two opinions. We consider 11 models, namely four open-weight judges from different families and one closed judge, together with the `Qwen2.5` family from 0.5B to 72B parameters. We run our experiments over 21,569 pairs coming from three datasets that include opinions regarding relevant political issues and moral dilemmas. We test LLM-as-a-Judge in three different scenarios that are used in practice: pairs of synthetic opinions, pairs of human opinions, and pairs of one human and one synthetic opinion, as in reward-based training. For each pair of opinions we consider five prompt variants and two directions, for a total of 215,690 judgements per model.

We focus on two properties that a reliable judge should satisfy: the score assigned to a pair should not change when the scoring instruction is rephrased, nor when the two opinions are swapped. Specifically, we address the following research questions:

**RQ1** How stable are LLM-as-a-Judge scores of opinion alignment when the scoring instruction is rephrased or the two opinions are swapped?

**RQ2** Which properties of the opinions are associated with unstable scores, and are the same pairs unstable across judges?

**RQ3** When the score depends on the order of the two opinions, is the effect systematic, i.e. does the LLM-as-a-Judge consistently reward or penalise opinions with a given property depending on whether they act as the reference or as the candidate?

**RQ4** How do stability and directional bias change with model scale?

Our contributions are twofold. First, we propose a framework for assessing the reliability of LLM-as-a-Judge in opinion alignment without human ratings, based on invariance to the phrasing of the instruction and to the order of the two opinions. We adapt two metrics from Shi et al. [23], Repetition Stability and Position Consistency, to the alignment setting and introduce a directional bias analysis that tests whether

[1] Language models are known not to represent human perspectives uniformly and may systematically favour some viewpoints over others [17, 18]. In particular, previous work has shown that LLMs tend to align with the positions of WEIRD (Western, Educated, Industrialised, Rich and Democratic) populations [18, 19].

order effects consistently assign higher rewards to opinions with a given property. Second, we conduct a comprehensive empirical study of these properties across four open-weight judges, one closed judge, and seven sizes of the `Qwen2.5` family to test the effect of scale on stability, using synthetic, human, and mixed human-synthetic opinions. We further examine how instability and order effects relate to properties of the opinions, including their length, sentiment, and semantic similarity, as well as the judge's perplexity and stance on the situation.

Additionally, we release our code and data[2], including the generated synthetic opinions and the scores assigned by all eleven judges across prompt variants and opinion orders, to support further research on judge reliability in opinion alignment.

Our experiments reveal several issues which should be taken into account when using LLM-as-a-Judge approaches for opinion alignment evaluation and rewards:

- **Judges are unstable, even though they appear stable on average.** Every judge changes its score by more than **30**% of the scoring scale when the opinions are swapped in at least **10**% of pairs. This instability is not random: a subset of pairs makes all judges unstable far more often than expected by chance, and instability is associated with differences in length, sentiment and semantic similarity between the two opinions, and with whether the judge itself agrees more with the stance of one of them. In practice, judges are least reliable on the pairs that matter most for opinion alignment: when the generated opinion differs in length from the reference, when it is only partly similar to it, and when it takes a different stance on a situation where the judge leans towards one side.
- **Order effects are systematic**. Almost every judge rewards candidate texts that are longer than the reference, while open-weight judges reward candidates that share their own stance on certain situations. As a result, a judge may prefer a longer candidate, or one that agrees with its own stance, over a candidate that more accurately reproduces the reference opinion. When such scores are used as rewards, these preferences become training signals, steering models towards more verbose outputs and a narrower range of viewpoints.
- **Larger judges are more stable, but not necessarily less biased.** As model size increases, there are fewer instances where the model's scores change when the two opinions are swapped. However, the instability that remains in larger judges is more consistent in direction. The score shifts tend to favour candidates with similar characteristics rather than varying randomly across pairs. This turns instability from noise into systematic bias.

The rest of this paper is organised as follows. Section 2 reviews the literature on LLM-as-a-Judge and its biases, and on the use of LLM-as-a-Judge for opinion alignment. Section 3 presents the experimental setup and the methodology of the stability and directional bias analyses, together with the features we examine and the datasets and models we use. Section 4 reports the results of both analyses on judges from different model families, followed by a post-hoc analysis of the situations and pairs that lead to the highest instability, and an analysis of the effect of model scale on

[2] https://github.com/SDavenia/opinionjudge

judges of different sizes from the `Qwen` family. Finally, Section 5 summarises our main findings, discusses the limitations of this study and outlines directions for future work.

## 2 Related Works

Since the emergence of large language models, their ability to perform a wide range of tasks through natural-language prompting has enabled their application to problems traditionally addressed by task-specific NLP systems [24, 25]. One frequent use-case scenario is LLM-as-a-Judge, where the model is used to score, choose or judge instances of textual content in a variety of methodological settings [4–6, 26]. This framework is largely adopted for the evaluations of generated text, as it combines the advantages of automatic evaluation, lower costs and shorter times compared to human evaluations, but goes beyond shallow automatic metrics such as BLEU, ROUGE and others [1–3]. Despite its popularity, LLM-as-a-Judge has proven to be imperfect, with a large number of works studying its biases and unreliability [8, 23, 27]. As our work focuses on evaluating the potential biases of LLM-as-a-Judge in the context of opinions, we first discuss the literature regarding the evaluation of LLM-as-a-Judge at large, and then focus on the employment of LLM-as-a-Judge in opinion-alignment frameworks. To the best of our knowledge this is the first work that studies LLM biases in this framework, which is highly relevant given the subjective nature of the topic.

### 2.1 LLM-as-a-Judge Biases

Despite its advantages, LLM-as-a-Judge can exhibit systematic biases that affect the reliability of its evaluations. Several studies have investigated these biases, focusing on model preferences, positional effects and the consistency of judgements across different evaluation conditions.

A first line of research investigates self-preference bias, whereby LLM judges tend to favour responses generated by themselves or by models with similar characteristics. Wataoka et al. [27] study self-preference in pairwise evaluation settings, in which a judge model must select the better response among two candidates given a specific scenario. Their experiments rely on the Chatbot Arena dataset [4], a crowdsourced collection of approximately 33,000 open-ended dialogues, each consisting of a user query, two responses produced by different LLMs, and a human pairwise preference label. Their findings indicate that LLMs tend to prefer their own generations over those produced by other models, as assessed against human judgments. They also identify a relationship between response perplexity and evaluation preferences, showing that judges tend to favour responses for which they assign lower perplexity. These findings suggest that evaluation outcomes may depend not only on the quality of the evaluated responses but also on their relationship with the judge model. Building on this, Panickssery et al. [28] investigate whether LLM evaluators favour and recognize their own generations on the task of news summarization, indicating that self-recognition is associated with self-preference. Together these studies highlight the importance of distinguishing preferences arising from the characteristics of the judge from judgements reflecting the actual quality of the evaluated content.

Beyond self-preference, researchers have investigated the effects of input characteristics on evaluation outcomes. Ye et al. [8] introduce CALM, a framework for systematically examining biases in LLM-based evaluation through perturbations of the evaluated inputs. The framework considers several forms of bias, including position, verbosity, diversity, and self-enhancement bias. To this end, the authors constructed three datasets from existing resources: an alignment dataset sampled from DPO preference-pair datasets, a fact-related dataset of mathematical and scientific reasoning questions and a refinement-aware dataset of general-knowledge questions. By examining how judgments change under different perturbations, the authors assess the extent to which evaluation outcomes are sensitive to specific characteristics of the input. This approach highlights the importance of evaluating judges not only in terms of their agreement with reference judgments but also in terms of their robustness to changes that should not affect the underlying evaluation.

Position bias has also been explicitly addressed by Li et al. [7], who investigate the tendency of pairwise LLM evaluators to favour candidates depending on their position in the input. They propose PORTIA, an approach that splits candidate responses into segments and aligns them before merging them into an evaluation prompt, which improves evaluation consistency. Their experiments are conducted on MT-Bench [4], an instruction-following benchmark of 80 open-ended questions spanning eight categories (e.g., writing, reasoning, coding, and humanities), paired with answers generated by different LLMs.

While these studies investigate particular sources of bias, Shi et al. [23] examine the reliability of LLM judges from the complementary perspectives of variability and position consistency. They investigate whether judgments remain stable when the same query is evaluated repeatedly and whether the preferred candidate changes when the order of the alternatives is permuted. Their analysis covers two benchmarks: MT-Bench [4], consisting of open-ended general-purpose questions, and DevBench [29], a software-development benchmark, in both cases comparing candidate model outputs against a fixed baseline. The authors formalize metrics for evaluating judgment reliability, position consistency, and primacy and recency effects. Their work highlights the distinction between the stability of repeated judgments and consistency across different candidate orderings. We adapt their repetition stability and position consistency metrics to our setting, as described in Section 3.2.

Taken together, these studies demonstrate that the reliability of LLM-as-a-Judge can be affected by the characteristics of the judge model, the properties and presentation of the evaluated responses, and the conditions under which judgments are elicited. Notably, the datasets employed in these works cover open-domain dialogue, summarization, mathematical and factual question answering, and software development, but none of them targets opinions. As a result, the implications of these biases for opinion-alignment evaluation remain insufficiently understood. In particular, when the evaluation concerns opinions, it is important to determine whether judgments reflect the opinions expressed in the evaluated content or are influenced by other characteristics of the responses or the evaluation process.

Moreover, these biases have been studied in settings where the judge is asked to select the best response among two or more candidates, to rank them, or to assign each

of them an absolute quality score. In all these cases, the judge is evaluating *quality*: there exists, at least in principle, a better and a worse answer, and human preferences or reference judgments serve as ground truth for that ordering. The alignment setting adopted in work on aligning LLMs to human behaviours, described below, differs in a fundamental way. Here, the judge is not asked which response is better, but how closely a generated response matches a reference response produced by a human, for instance whether it expresses the same stance and relies on the same reasoning. The evaluated quantity is therefore *similarity to a target* rather than quality, and the two texts play distinct roles, a reference and a candidate, rather than being interchangeable alternatives competing for preference. Consequently, it is unclear to what extent findings on position, verbosity, or self-preference bias transfer to this setting.

## 2.2 LLM-as-a-Judge for Opinion-Alignment

LLM-as-a-Judge can be employed in settings where the judge is asked to assess how well a model-generated answer reflects a human-generated gold answer. This setting has become central to work on simulating human behaviour with LLMs, where the goal is not to produce the best possible response, but the response a person would give. The evaluation metric produced by the judge can be used as an evaluation tool to quantify model performance, or, more crucially, as a reward signal for reinforcement learning.

Wu et al. [12] propose HumanLM, a framework for training user simulators that align with the latent states of real users rather than imitating their surface language. They introduce Humanual, a benchmark of six datasets built from real user responses, including opinions on personal and moral dilemmas from Reddit (Humanual-Opinion) and responses to political blog posts (Humanual-Politics). An LLM judge assigns an alignment score in $[0, 1]$ measuring how consistent a generated response is with the ground-truth response. The judge serves as the reward during training (GPT-5-mini) and as the evaluation metric at test time (claude-4.5-haiku).

A second group of works uses the judge to compare simulated and real users along stylistic and behavioural dimensions. MirrorBench [16] evaluates the human-likeness of user-proxy agents. The judge is presented with a proxy-generated conversation and the corresponding human reference, and returns a similarity score in $[0, 1]$ focused on aspects such as tone and naturalness. The authors explicitly acknowledge that judges may favour verbosity, politeness, or their own outputs. They report that absolute scores, and sometimes model rankings, shift with the choice of judge. Similarly, SimulatorArena [15] collects real human–LLM conversations on math tutoring and document creation, the latter referred to as SimArena-Doc in subsequent work. A GPT-4o judge rates the similarity between simulated and real user messages on a 1–5 Likert scale, separately for writing style and interaction style.

CoSER [14] addresses role-playing of established characters, using authentic multi-character dialogues extracted from novels. An actor model re-enacts each scene, and an LLM critic (GPT-4o) evaluates the simulated conversation against the original dialogue along four dimensions: storyline consistency, anthropomorphism, character fidelity, and storyline quality. In its original formulation, scoring is penalty-based: the critic lists flaws with an associated severity, which are subtracted from a maximum

score. A length correction compensates for the observation that longer simulations accumulate more flaws.

Finally, OdysSim [13] unifies 23 human-behaviour simulation tasks, including Humanual, MirrorBench, SimArena-Doc, and CoSER, into a single benchmark with scores normalised to $[0, 1]$. LLM judges serve both as the evaluation metric and as the reinforcement-learning environment, returning a scalar reward together with a textual critique (GPT-5-nano by default, GPT-5.4 for CoSER). OdysSim also reports reward-hacking patterns induced by judge-based rewards. In some tasks, the policy learns to insert evaluation-like statements that target the judge directly. In Humanual and CoSER, error-counting rewards lead the policy to produce short, generic responses that contain fewer checkable claims. To mitigate the latter, the authors replace error counting with rubric-based scoring and add length-matching and lexical-overlap rewards against the ground truth.

These works treat the judge as a reliable proxy for alignment with a human reference, and its scores directly determine both model rankings and training signals. Each work introduces ad hoc safeguards against specific failure modes: different judges for training and testing, comparative scoring, calibration controls, order swapping, length correction, and reward-hacking detectors. However, none systematically examines the stability of the judge or its biases in this setting. This gap is particularly relevant when the reference expresses an opinion, as in Humanual-Opinion and Humanual-Politics. In such cases, alignment depends on the stance and reasoning conveyed by the two texts, and a judge whose scores are driven by its own preferences or by surface features of the responses would systematically reward some opinions over others. Our work addresses this question directly in the context of opinion alignment.

# 3 Methodology

For the scope of this work, we systematically evaluate various models acting as LLM-as-a-Judge for the alignment of pairs of opinions $x$ and $x'$ on situation $s$, assessing their robustness and reliability. Instead of directly relying on human ratings, we assume that reliable judges should assign consistent scores to semantically equivalent settings. In particular, we focus on instances that differ only in the order in which the opinions are presented or in the wording of the prompt[3].

We study two complementary aspects of LLM-as-a-Judge's reliability: how *stable* its scores are across such variations (*stability analysis*), and whether this change is directional, i.e., whether presenting an opinion with a given characteristic first leads to consistently higher or lower scores (*directional bias analysis*).

Specifically, the *stability analysis* aims to identify which features of the input, e.g., properties of $x$, $x'$, or $s$,[4] are most predictive of judge instability, where instability is measured as the variation in scores across prompt variants or across the two orderings of the pair. Beyond characterizing LLM-as-a-Judge's reliability, this analysis is intended as a practical guide for researchers using LLM-as-a-Judge for evaluation or alignment in NLP tasks working with opinions, indicating which input conditions are

[3] We verify and further discuss this assumption by comparing models to human annotators (Appendix A.1): human annotators are more stable than models when scoring pairs in reverse order.

[4] A feature could be for example the text length of $x, x'$ or the model's perplexity on $x, x'$.

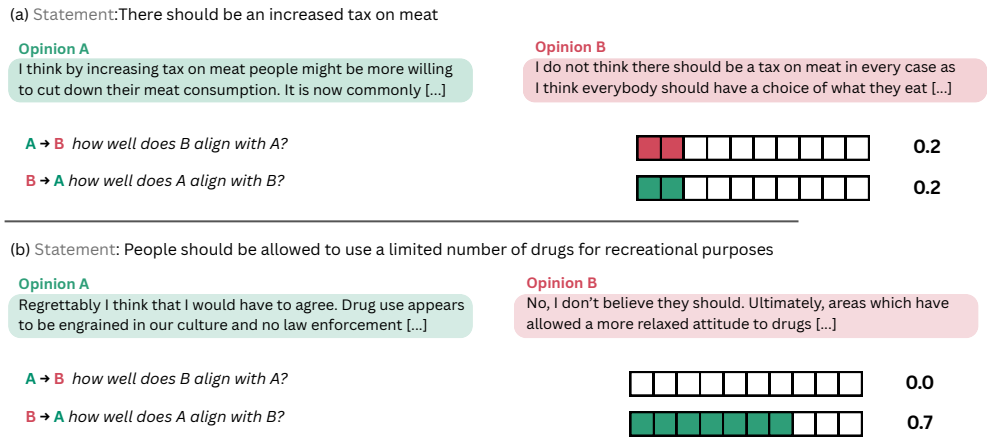



**Fig. 2**: Scoring on two Habermas examples obtained with `Llama-70B` as a judge. On the first example the judge is stable, while on the second the change in score is very large.

likely to yield more or less robust scores. Figure 2 shows two examples of scoring from our stability analysis, where (a) is a stable judgement with respect to positionality, while (b) results in a high score change when opinions are swapped.

The *directional bias analysis* instead asks whether the score is systematically, rather than randomly, affected by position: that is, whether presenting an opinion with a given characteristic first (i.e., as the reference) versus second (i.e., as the candidate) in the pair consistently inflates or deflates the resulting score. Such a directional bias would be consequential in two ways: as a training signal, it could introduce spurious, order-dependent variation in the reward or penalty assigned to otherwise-equivalent generations; as an evaluation metric, it could make reported scores systematically too lenient or too conservative depending on an essentially arbitrary convention.

First, we introduce a formalization of the setting of our experiments (Section 3.1). In Section 3.2, we introduce the metrics used to evaluate the stability of the judges and, in Section 3.3, we explain the methodology used to analyse the presence of consistent directional biases. Afterwards, in Section 3.4 we discuss the dimensions (textual- and model dependent-characteristics) that we selected for our analysis. Finally, we report the data sources used for evaluating the judges (Section 3.5) and the selected models in Section 3.6.

## 3.1 Judge Setting

Each LLM-as-a-Judge is evaluated on the task of scoring the similarity between two opinions related to a specific situation. The judge has to score each pair bidirectionally and over multiple prompts.

Formally, for each dataset we have a set of situations $\{c_s\}_{s=1}^{S}$, each with a set of related textual opinions $\{x_{si}\}_{i=1}^{N_s}$. We consider all pairs of opinions $(x_{si}, x_{sj})$ on the same situation, with $1 \leq i < j \leq N_s$. To test the stability of the judge to

prompt variability, we score each pair under a set $\{p_k\}_{k=1}^{K}$ of $K$ semantically equivalent prompt variants (see Appendix B; in our experiments $K = 5$), and we denote by $y_{sijk} = \mathrm{LLM}(x_{si}, x_{sj} \mid c_s, p_k)$ the score obtained when $x_{si}$ is presented first (i.e. "How well opinion $x_{sj}$ reflects opinion $x_{si}$"). Since alignment between two opinions should not depend on the order in which they are shown, we also score every pair with the opinions swapped, $y'_{sijk} = \mathrm{LLM}(x_{sj}, x_{si} \mid c_s, p_k)$ (i.e. "How well opinion $x_{si}$ reflects opinion $x_{sj}$").

Following previous work presented in Section 2, we formalize judge ratings on a continuous scale in $[0, 1]$, with 1 indicating perfect alignment between the two opinions and 0 total misalignment.

## 3.2 Stability Analysis

We first ask whether the scores of LLM-as-a-Judge are stable across semantically equivalent queries. Rephrasing the instruction or swapping the two opinions should not influence the scoring. To quantify stability, we adapt two metrics from Shi et al. [23].

Repetition Stability ($RS$) measures sensitivity to how the instruction is phrased, and Position Consistency ($PC$) measures sensitivity to the order of the two opinions. The original metrics are defined for pairwise choice: the judge selects the better of two responses, and stability is measured as agreement between these discrete choices. In our setting, LLM-as-a-Judge returns a continuous score in $[0, 1]$ describing the relation between two opinions, so we replace agreement between choices with absolute differences between scores.

Because both metrics are defined at the level of individual pairs, they also let us examine whether instability is concentrated in pairs with specific textual characteristics, such as a large difference in length between the two opinions. Identifying these characteristics indicates which properties of the opinions should be controlled for when using LLM-as-a-Judge.

#### *Repetition Stability ($RS$)*

Repetition Stability ($RS$) asks whether the LLM-as-a-Judge assigns the same score to a pair of opinions when the instruction is rephrased. Rather than repeating an identical query under sampling-based decoding and averaging over a fixed number of trials, as in Shi et al. [23], who sample 3 times from a single prompt, we compute $RS$ over semantically equivalent prompt variants, all generated under greedy decoding. We argue this better reflects how LLM-as-a-Judge is deployed in practice: judges are typically queried with greedy rather than stochastic decoding [30], so sampling variance is seldom encountered in real use, whereas variation in how the scoring task is phrased is common across implementations, papers, and practitioners, and, to our knowledge, unexamined. The metric therefore captures sensitivity to how the task is phrased, rather than to sampling randomness.

For the pair of opinions $(x_{si}, x_{sj})$ of situation $s$, we measure how much its scores in a fixed order vary across the $K$ prompt variations through:

$$d_{sij} = \frac{1}{D} \sum_{1 \le k < l \le K} |y_{sijk} - y_{sijl}|, \tag{1}$$

where $D$ is the largest value the sum can take for scores in $[0, 1]$ (reached when the scores are split as evenly as possible between 0 and 1), so that $d_{sij} \in [0, 1]$. We define $d'_{sij}$ in the same way on the reversed-order scores $y'_{sijk}$. The repetition stability of the pair is then

$$RS_{sij} = 1 - \frac{1}{2}\big(d_{sij} + d'_{sij}\big). \tag{2}$$

The two orders are treated separately and only then averaged, so that the effect of swapping the opinions is captured by $PC$ and not here. $RS_{sij} = 1$ means that the LLM-as-a-Judge assigns the pair the same score under every prompt variation. The overall repetition stability of a LLM-as-a-Judge is the average over all pairs,

$$RS = \frac{1}{M} \sum_{s=1}^{S} \sum_{1 \le i < j \le N_s} RS_{sij}, \tag{3}$$

where $M$ is the total number of pairs across all situations and $N_s$ is the total number of opinions per situation $s$. Lower values indicate that the LLM-as-a-Judge assigns substantially different scores to the same pair under semantically equivalent prompts.[5]

***Position Consistency** (PC)*

Position consistency instead asks whether the LLM-as-a-Judge assigns the same score to the questions "How well does opinion B align with opinion A?" and vice-versa, under the symmetric assumption.

For each prompt variation, the score of a pair in one order is compared with its score in the reversed order. For $(x_{si}, x_{sj})$, the metric at the pair level is defined as:

$$PC_{sij} = 1 - \frac{1}{K} \sum_{k=1}^{K} |y_{sijk} - y'_{sijk}|, \tag{4}$$

and the overall $PC$ is again the average over all pairs,

$$PC = \frac{1}{M} \sum_{s=1}^{S} \sum_{1 \le i < j \le N_s} PC_{sij}. \tag{5}$$

[5] $RS$ ranges from 0 to 1, where 1 signifies perfect stability under prompt variations. A LLM-as-a-Judge that assigns scores uniformly at random has an expected $RS$ of about 0.44 for $K = 5$. Since scores are bounded in $[0, 1]$, how much they can vary also depends on where they lie on the scale: a score of 0.5 can differ from another score by at most 0.5, whereas a score of 0 or 1 can differ by up to 1. Pairs that receive intermediate scores therefore have less room for instability.

$PC_{sij} = 1$ means that, under every prompt variation, changing the order of the opinions does not affect the judgement of the LLM-as-a-Judge.[6]

### 3.3 Directional Bias Analysis

The metrics introduced in Section 3.2 measure whether swapping the two opinions changes the score, but not whether the change has a systematic direction. For instance, a LLM-as-a-Judge may consistently give higher scores when the longer opinion is presented last (i.e., as a candidate answer). To study this, for each feature we select one opinion of the pair, the *pivot*, through a feature-specific rule (listed below), and ask whether the LLM-as-a-Judge scores the pair higher when the pivot is presented first (*pivot-first*) or when it is presented second (*pivot-last* preference).

Taking $x_{si}$ to be the pivot of the pair $(x_{si}, x_{sj})$, we compute for each prompt variation the oriented score difference $\delta_{sijk} = y_{sijk} - y'_{sijk}$, i.e. the score obtained when the pivot is presented first (as the reference), minus the score obtained when it is presented second (as the candidate). Averaging over prompt variations gives the pair-level difference

$$\Delta_{sij} = \frac{1}{K} \sum_{k=1}^{K} \delta_{sijk}. \tag{6}$$

If the feature has no effect on the LLM-as-a-Judge's positional behaviour, $\Delta_{sij}$ is centred on zero. Positive values indicate a *pivot-first* preference, and negative values a *pivot-last* preference. Note that **"preference" here does not refer to a choice between the two opinions, as the judge never selects one opinion over the other**. Instead, the score is assigned at the pair-level, namely to the questions "How well the candidate $x'$ aligns with the reference $x$", and vice-versa. Within this setup, a *pivot-last* preference means that the judge assigns a higher score to the pair when the pivot opinion is presented as the candidate rather than the reverse.

$\Delta_{sij}$ measures how large the asymmetry is, but not how consistent it is: a few large differences can produce the same mean as many small ones. For a set of pairs (e.g. all pairs within a range of the feature gap), we therefore also compute the directional consistency

$$C = \frac{1}{|\mathcal{P}|K} \sum_{(s,i,j)\in\mathcal{P}} \sum_{k=1}^{K} \operatorname{sign}\big(\delta_{sijk}\big) \in [-1, 1], \tag{7}$$

the share of judgements in which the pivot is scored higher when presented first, minus the share in which it is scored lower. $C = 1$ means that the pivot always scores higher when presented first, $C = -1$ that it always scores lower, and $C = 0$ that there is no preferred direction (all ties, or equally many differences in each direction).

The direction of the effect matters because the two positions play different roles when the judge is used to align or evaluate a model to a human reference: the first position contains the reference and the second contains the generated candidate.

[6] $PC$ ranges from 0 to 1, where 1 signifies perfect consistency under order changes. A LLM-as-a-Judge that assigns scores uniformly at random has an expected $PC$ of about 0.67. As for $RS$, the room for inconsistency depends on where the scores lie on the scale: if a pair receives a score of 0.5 in one order, its $PC$ for that prompt variation cannot fall below 0.5.

Suppose text length is the feature under analysis, and the longer text is the pivot. If the judge assigns higher similarity when the longer text comes first (*pivot-first* preference), candidates shorter than the reference will be systematically rewarded over long ones. Conversely, if it assigns higher similarity when the longer text comes second (*pivot-last* preference), candidates longer than the reference will be rewarded over short ones. In both cases, the policy can exploit the bias by drifting in length rather than by better matching the reference opinion and the measured alignment is confounded with some form of proximity to what the LLM-as-a-Judge prefers.

## 3.4 Analysis Features

To characterize which text or judges characteristics are linked to inconsistencies in the evaluation, we select a set of features, starting from forms of biases which are already studied in other LLM-as-a-Judge settings, and complement them with opinion-specific features.

We consider both features that depend on the text only (e.g., text length) and features that describe how the model represents and evaluates the text (e.g., model's perplexity).

Some features are extracted from the situation (e.g., whether for a given situation the model prefers opinions that support or oppose it), while others are extracted from the opinions themselves (e.g., text length is specific to each opinion).

We describe these features in more detail below, and we report whether they are constructed from the situation (SIT) or opinions (OP).

***Text-only features***

These dimensions are computed directly from the text itself (the situation or the textual opinion), independent of any judge model. They capture intrinsic properties of the content, namely how long it is, what emotional tone it carries, how safe it is flagged to be, that could plausibly correlate with judge instability or bias. For each dimension we also report the pivot used for the Directional Bias analysis.

- **Length (OP)**: The surface length of the opinion measured in number of words This targets *verbosity bias*, a well-documented tendency of LLM judges to favor longer responses regardless of their actual quality or correctness [8]. Each opinion is assigned its text length counted as number of words. The longer opinion is used as the pivot.
- **Embedding (OP)**: The cosine similarity between the embeddings of the two opinion texts[7]. The objective of this analysis is to verify whether a model is more or less stable when it has to judge the relationship between texts of varying level of semantic similarity. Given the generation-pairwise nature of this feature, it is not included in the directional bias analysis.
- **Sentiment-based (OP)**: The sentiment polarity expressed in the opinion, to test whether *sentiment bias*, a preference for certain emotional framings over others, shapes judge scoring independent of content alignment. The choice of this feature

[7]The embeddings are extracted with Nomic Embed v2 [31]. As we use the embeddings for textual similarity, the model is prompted with the *STS* default setting which prepends "classification: " before the text to extract the embedding.

follows from previous work investigating how certain models exhibit preferences towards positive or negative emotions [8]. We considered different models to extract the sentiment of the texts:

  - **Roberta** We rely on `siebert/sentiment-roberta-large-english` [32] to extract the softmax probability over the positive class vs the negative one. For each text we have a score in [0, 1] indicating its positive sentiment. The pivot is the opinion with the most positive sentiment.
  - **VAD** We extract the VAD scores using `RobroKools/vad-bert`. For each text we have valence, arousal and dominance scores on the [1-5] scale, with higher scores indicating higher values for that feature. We analyse all three separately and for each specific dimension we select the opinion with the higher score as a pivot.
  - **VADER** We rely on VADER [33], a lexicon/rule-based sentiment analyser. VADER outputs a compound score in [-1, 1] where values above 0.05 indicate a positive sentence and values below -0.05 indicate a negative one. The pivot is the opinion with the higher compound score, i.e. the more positive one.

- **Safety guardrails (OP)**: The Guard-estimated probability that the *opinion* is unsafe. Each opinion is assigned a score in [0, 1], indicating the probability of being unsafe according to the Llama-Guard model. We rely on `meta-llama/Llama-Guard-3-8B` [34] to extract this probability by applying softmax to the logits of the *safe* and *unsafe* tokens. The goal of this analysis is to understand if LLM-as-a-Judge might behave differently in instances of texts that might potentially trigger safety mechanisms. We select as a pivot the opinion with the lower unsafe probability, i.e. the safer one.

***Features describing how the model represents the text***

Unlike the text-only dimensions above, these are not properties of the text in isolation, as they can only be computed *with respect to a specific judge model*, since they depend on how that model processes, or responds to the text. The same generation can yield a different perplexity, refusal probability, or alignment score depending on which model is used as the judge. These dimensions therefore capture the judge model's own implicit relationship to the text.

- **Perplexity-based (OP)**: The perplexity of the textual opinion under the judge model itself (conditioned on a neutral situation prompt, scoring only the opinion tokens). Lower perplexity indicates the text is more "fluent" or expected *to that particular judge model*; this is used as a proxy for whether the judge implicitly favours text that resembles its own generation distribution, following previous work in self-preference bias [27]. Since raw perplexity is heavily right-skewed and its scale differs across models, for visualization we report log-perplexity standardized (z-scored) within each judge model. The opinion with the lowest perplexity is used as the pivot.

For the two SIT features, we use a common notation for the three stances an opinion can take on a situation: supporting (S), neutral (N) and opposing (O), corresponding

to Supports, Either and Opposes in ValuePrism, and to Agree, Neutral and Disagree in Habermas (see Section 3.5 for full dataset details).

- **Refusal Probability (SIT)**: The probability that the judge model refuses to write an opinion taking a given stance on the situation. Unlike the text-only safety scores discussed above, this feature does not measure whether the *content* of an opinion is unsafe, but whether *this particular judge* is willing to argue for that stance at all. For each situation and stance, we prompt the judge to write an opinion taking that stance, sample 30 completions with temperature 1.0, and compute the fraction that are refusals, detected through the refusal-phrase matching of Arditi et al. [35].[8] The pivot is the opinion whose stance has the lower refusal probability. For example, with $P_{\text{ref}}(S) = 0.5$, $P_{\text{ref}}(N) = 0.1$ and $P_{\text{ref}}(O) = 0.2$, the pivot of a pair with an S and an O opinion is the O opinion.
- **Alignment (SIT)**: The stance that the judge model itself takes on the situation, measured as a probability distribution over S, N and O. To estimate it, we follow the approach of Scherrer et al. [36] and ask the judge for its own verdict on the situation, using 10 manually constructed, semantically equivalent prompts that require it to answer with one of three labels (e.g. ACCEPTABLE, UNACCEPTABLE or AMBIGUOUS) followed by a short rationale. For each prompt we sample 30 answers with temperature 1.0, map each label to the corresponding stance, and average the resulting frequencies over the ten prompts. Each opinion then takes the probability of its own stance: for example, if the judge answers S in 60% of cases, N in 30% and O in 10%, every supporting opinion on that situation has an alignment value of 0.6. The pivot is the opinion whose stance the judge is more likely to take: in the example above, the S opinion in a pair with an O opinion.

Note that pairs for which the rule does not identify a pivot are excluded from the directional analysis, i.e. pairs with equal feature values. For example, a pair whose two opinions have the same length is not included in the directional analysis for text length.

## 3.5 Data Sources

In this study, we consider three distinct types of pairs: fully synthetic pairs, fully human pairs, and mixed synthetic and human pairs. The fully synthetic dataset helps us analyse the phenomena in a more controlled manner; the fully human setting is valuable to interpret the use of LLM-as-a-Judge on noisier data (e.g., containing typos); lastly, the mixed human use-case gives us better insights on reward training-loops used for alignment.

***Fully Synthetic Pairs - ValuePrism***

To create a fully synthetic dataset of opinions that spans a broad range of topics and is varied in content, we leverage the metadata in ValuePrism (*VP*) [37]. VP is a collection of human-written moral situations (e.g., "Saving an animal by risking your life"). Each situation is paired with several values, rights and duties (the *vrd* field), each with

[8] To avoid false positives, matches are further verified with a classifier (`Llama-3.3-70B-Instruct`).

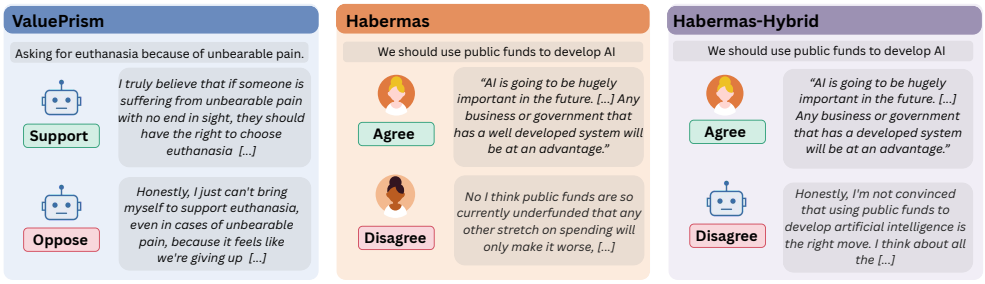



**Fig. 3**: Example instances from the ValuePrism (left), the Habermas (middle) and the Habermas-hybrid (right) datasets.

its own stance towards the situation (*Support*, *Oppose* or *Either*) and an *explanation* of how it relates to the situation. For the situation above, for instance, the *stance* and the *vrd* can be "support" and "courage", with the *explanation* "Taking risks to help others in need, including animals, can be seen as an expression of courage". As explanations are very concise and not in the style of a personal opinion, we generate synthetic textual opinions with `meta-llama/Llama-3.3-70B-Instruct` from the fields described above. More details on model and prompt selection in Appendix C. To ensure that the generated opinions expressed the information they were conditioned on, we conducted a quality annotation with three annotators checking 150 opinions sampled across target stances, and found that 97.3% of them were judged to keep target *stance*, *value* and *explanation*. The few failures are minor and none reverses the target stance. More details are provided in Appendix A.

### *Fully Human Pairs - Habermas*

The Habermas (*HB*) dataset [38] contains people's opinions on a set of political issues, using a sample of UK residents. Each situation is a policy statement answered by several participants. Each participant gives their own stance on it, on a 7-point scale that we collapse into *Agree*, *Neutral* and *Disagree*, and writes a free-text opinion.

### *Mixed Pairs - Habermas-hybrid*

To construct a dataset that mirrors the evaluation setup and/or reward-training loop of opinion-aligned models, we augment the *HB* dataset with synthetic opinions (*HBY*). For every sampled *HB* situation, we generate three synthetic opinions, one per stance (*Agree*, *Neutral*, *Disagree*). We construct pairs containing one human and one synthetic opinion, so each human-written opinion is compared with a synthetic opinion that shares its stance and with two that do not. This set-up is the closest to how LLM-as-a-Judge is used in practice, both in evaluation and as a training reward, where a generated opinion is scored against a human-written one. We therefore only score pairs made of one human and one synthetic opinion.

As the two original datasets are very large and our setting requires multiple judgements for each couple, we sample 500 situations with specific constraints from each corpus. In Appendix C, we give further details on the preprocessing and sampling from the datasets.

Examples from all three datasets are shown in Figure 3, in a standardised format which is used in the rest of the paper, where we only consider the situation, stance and opinion.

### 3.6 Models Evaluated

We evaluate a model from four popular families of open models: `meta-llama/Llama-3.3-70B-Instruct` (`Llama-70B`) [34], `google/gemma-3-27b-it` (`gemma-27B`) [39], `mistralai/Mistral-Small-3.2-24B-Instruct-2506` (`Mistral-24B`) [40], `Qwen/Qwen2.5-72B-Instruct` (`Qwen-72B`) [41]. All open models are used via the Transformers library and HuggingFace. Additionally, we also evaluate one closed model: `GPT-5.4-nano` via OpenRouter, which is selected as it is employed in the literature [13] and remains relatively cheap compared to other closed models.

In addition, to analyse the impact of model sizes, we test the entire `Qwen-2.5` family which has the most variety of models based on number of parameters. In particular we take into consideration `Qwen-0.5B`, `Qwen-1.5B`, `Qwen-3B`, `Qwen-7B`, `Qwen-14B`, `Qwen-32B` and `Qwen-72B` in their instruction variants.

All models are used in open generation setting with greedy decoding and a new token limit of 5. As the score of the model is a float, we parse it keeping only 2 decimal digits. We did not encounter any instance where the LLM-as-a-Judge failed to output a valid score.

Refusal probability and alignment are computed only for the open judges, since they require hundreds of sampled generations per situation, which would be costly through a commercial API. Perplexity is likewise restricted to the open judges, because it requires the log-probabilities of the opinion tokens, which `GPT-5.4-nano` does not expose.

## 4 Results

This Section presents the stability (Section 4.1) and directional bias (Section 4.2) analyses. We complement them with an analysis of whether the most unstable pairs and situations are shared across judges (Section 4.3), and conclude by analysing the effect of model size on stability and bias using the `Qwen` family (Section 4.4).

As the analyses share many common components, we summarize their key units below.

***Pair-level metrics.***

The unit of analysis is a pair of opinions $(x_{si}, x_{sj})$. For the stability analysis, its metrics are the repetition stability $RS_{sij}$ (capturing score variation across different prompts) and the position consistency $PC_{sij}$ (capturing score variation when the opinions are

swapped). For the directional bias analysis, the metric is the oriented score difference $\Delta_{sij}$ (measuring whether a model exhibits *pivot-first* or *pivot-last* preferences on each specific feature).

***Pair-level features.***

Every feature except embedding similarity assigns a value $f(x_{si})$ to each opinion. For the features computed on the opinion texts (OP), this is a property of the opinion itself, e.g. its length in words or its perplexity under the judge. The features computed on the situation (SIT) are instead estimated once for each situation and stance, and each opinion takes the value of its own stance: for alignment, $f(x_{si})$ is the probability that the judge itself takes the stance of $x_{si}$ on situation $s$, and for refusal, the probability that the judge refuses to voice an opinion with that stance. Two opinions with the same stance on the same situation therefore receive the same value. Both analyses characterise a pair by the gap between its two opinions:[9]

$$g_{sij} = |f(x_{si}) - f(x_{sj})|$$

The gap is symmetric, so in the stability analysis it does not depend on the order of the two opinions. In the directional bias analysis, the order is fixed by the pivot, and pairs whose two opinions have the same value have no pivot and are excluded; for alignment and refusal, this includes all pairs of opinions with the same stance.

***Statistical significance.***

For every feature and judge, we compute the Spearman correlation $\rho$ between the feature gap $g_{sij}$ (for embedding similarity, the similarity itself) and the pair-level metrics, with 95% confidence intervals and two-sided p-values for $\rho = 0$ obtained by bootstrapping.[10] Since with this many pairs even negligible correlations can be statistically significant [43], a relationship is marked with $*$ only if its adjusted p-value is below 0.05 and $|\rho| \geq 0.1$.

***Correlated features.***

Some features are correlated: the sentiment and affect scores measure related properties of the text, and longer opinions tend to have lower perplexity. To check that the associations we report are not explained by a correlated feature, we also model all features jointly, using partial correlations for stability and a joint linear model of $\Delta_{sij}$ for directional bias. The vast majority of associations remain significant with the same sign. The few that lose significance are mostly sentiment or affect features: since these features capture related aspects of the same property, their shared effect is often attributed to just one of them in the joint model. We therefore interpret the sentiment and affect features as a group rather than individually. Full results for the joint modelling are reported in Appendix D.4.

[9] For the stability analysis, we also consider the mean, maximum and minimum of the two values; these results are reported in Appendix D.

[10] Pairs from the same situation are not independent, so we resample at the situation level: in each of the 1,000 bootstrap samples, every resampled situation contributes all of its pairs. P-values are corrected with Benjamini–Hochberg [42] at a 5% false discovery rate, separately for each feature and dataset, jointly over all judges and, for stability, both metrics and the four aggregations of the feature.

| Dataset | Metric | Gemma-27B | Llama-70B | Mistral-24B | Qwen-72B | GPT-5.4-nano |
|---|---|---|---|---|---|---|
| *VP* | *RS* | $0.906_{\pm 0.061}$ | $\mathbf{0.928}_{\pm 0.071}$ | $0.895_{\pm 0.071}$ | $0.889_{\pm 0.081}$ | $0.811_{\pm 0.123}$ |
| | *PC* | $0.907_{\pm 0.096}$ | $0.912_{\pm 0.107}$ | $\mathbf{0.916}_{\pm 0.074}$ | $0.908_{\pm 0.090}$ | $0.856_{\pm 0.112}$ |
| *HB* | *RS* | $\mathbf{0.902}_{\pm 0.068}$ | $0.900_{\pm 0.102}$ | $0.891_{\pm 0.067}$ | $0.892_{\pm 0.079}$ | $0.773_{\pm 0.135}$ |
| | *PC* | $0.894_{\pm 0.104}$ | $0.889_{\pm 0.127}$ | $\mathbf{0.900}_{\pm 0.086}$ | $0.896_{\pm 0.097}$ | $0.841_{\pm 0.119}$ |
| *HBY* | *RS* | $0.902_{\pm 0.066}$ | $\mathbf{0.904}_{\pm 0.102}$ | $0.896_{\pm 0.065}$ | $0.892_{\pm 0.080}$ | $0.783_{\pm 0.142}$ |
| | *PC* | $0.890_{\pm 0.106}$ | $0.894_{\pm 0.117}$ | $0.888_{\pm 0.094}$ | $\mathbf{0.899}_{\pm 0.094}$ | $0.839_{\pm 0.130}$ |

**Table 1**: Stability results on the *VP*, *HB* and *HBY* datasets. *RS* and *PC* denote the two evaluation metrics. Subscripts give the standard deviation across opinion pairs of the pair-level *RS*/*PC*, each computed over that pair's repeated judgments. Best results are shown in bold.

***Interpretation of the results.***

We briefly explain how to read the results in the following sections. The metrics' scales theoretically range from 0 to 1 but their values need to be interpreted with the average judge scores in mind.

**Stability analysis**: *RS* and *PC* both range from 0 to 1, with 1 indicating perfect stability.[11] Note that a *PC* of 0.9 means that swapping the two opinions changes the score by 0.1 on average. Since the average score of the judges lies between 0.35 and 0.55 (Appendix D.1), this corresponds to roughly 18–29% of a typical score, which is consequential for real use-cases.

**Directional Bias Analysis**: Similarly, the directional score difference $\overline{\Delta}$ theoretically ranges from $-1$ to 1, but a bias in absolute value as small as 0.04 corresponds to roughly 7–11% of a typical score. Moreover, unlike random noise, this shift is systematic as it consistently favours the same opinion in a pair, so it does not average out over many pairs and influences reward-based alignment.

## 4.1 Stability Analysis

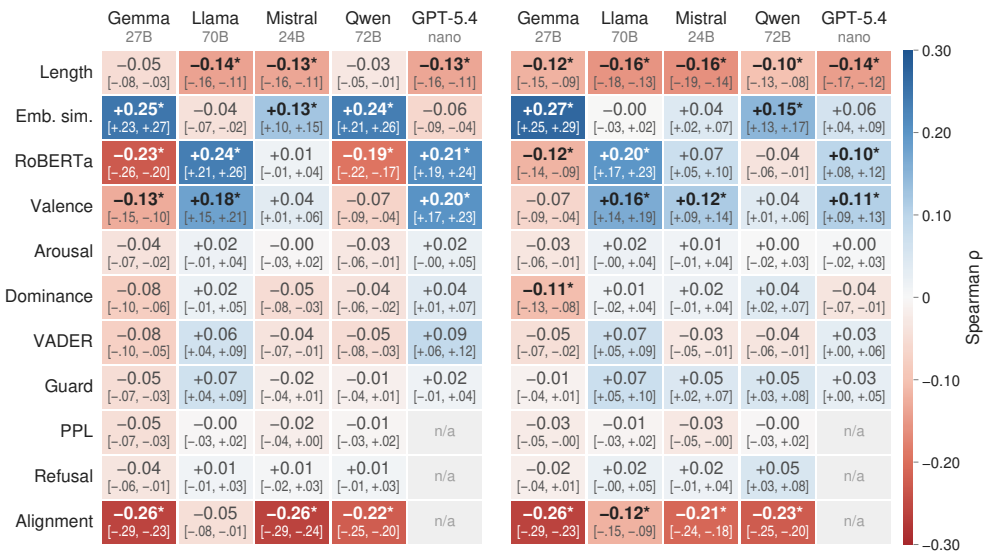
In this section, we first present an overview of the stability of each judge on the three datasets in terms of *RS* and *PC*. Then, we analyse how these metrics relate to the features.

***Overall results***

Table 1 reports the *RS* and *PC* obtained by each model on the three datasets.

The four open models perform within one standard deviation of each other, regardless of size (`Gemma-27B` and `Mistral-24B` are in the 20B range, `Llama-70B` and `Qwen-72B` in the 70B range). The closed model, `GPT-5.4-nano`, is less stable than all of them, especially in *RS*: it trails the open models by 0.08–0.13 in *RS* and by 0.05–0.06 in *PC*.

[11] However, a judge that assigns uniformly random scores in $[0, 1]$ would obtain an expected *RS* of 0.44 and an expected *PC* of 0.67.




(a) $RS_{sij}$ (b) $PC_{sij}$

**Fig. 4**: *VP*: Spearman $\rho$ [95% CI] between $RS_{sij}$ / $PC_{sij}$ and the difference in feature between the two instances. $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$.

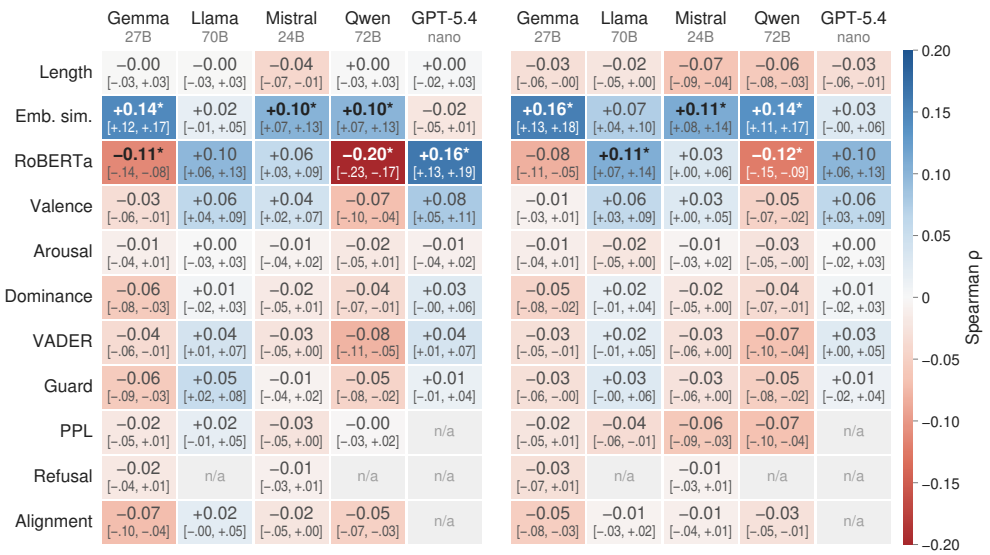
For the open models, *RS* and *PC* are close, differing by at most 0.021 for any judge and dataset, and neither is consistently higher. `GPT-5.4-nano` is the exception, with *PC* 0.05–0.07 above *RS*.

Stability does not change by much across datasets. Almost every score is slightly lower on *HB* than on *VP* (the only exception is the *RS* of `Qwen-72B`). *HBY*, where each pair consists of a human-written opinion and a model-generated one, stays close to *HB*: no score differs by more than 0.012 between the two.

Figures 4, 5 and 6 show, for each judge and feature, the Spearman correlation between the difference in feature between the two opinions ($g_{sij}$) and $RS_{sij}$ (left) or $PC_{sij}$ (right), on *VP*, *HB* and *HBY* respectively. Figure 7 complements these heatmaps by showing how $RS_{sij}$ and $PC_{sij}$ vary as the gap grows, for a selection of the most significant features; while the others are reported in Appendix D.

Overall, all significant correlations are small ($|\rho| \leq 0.27$), so no single feature explains a large share of the instability. Still, a few consistent patterns emerge. Four features account for most of the significant effects: text length, embedding similarity, sentiment and alignment. Only embedding similarity and sentiment are significant on all three datasets, while text length and alignment matter on *VP* and *HBY* but not on *HB*.

The two metrics are not necessarily affected in the same way: on *HBY*, for instance, a larger length or perplexity gap lowers $PC_{sij}$ but leaves $RS_{sij}$ unchanged (Figures 7b and 7e).




(a) $RS_{sij}$ (b) $PC_{sij}$

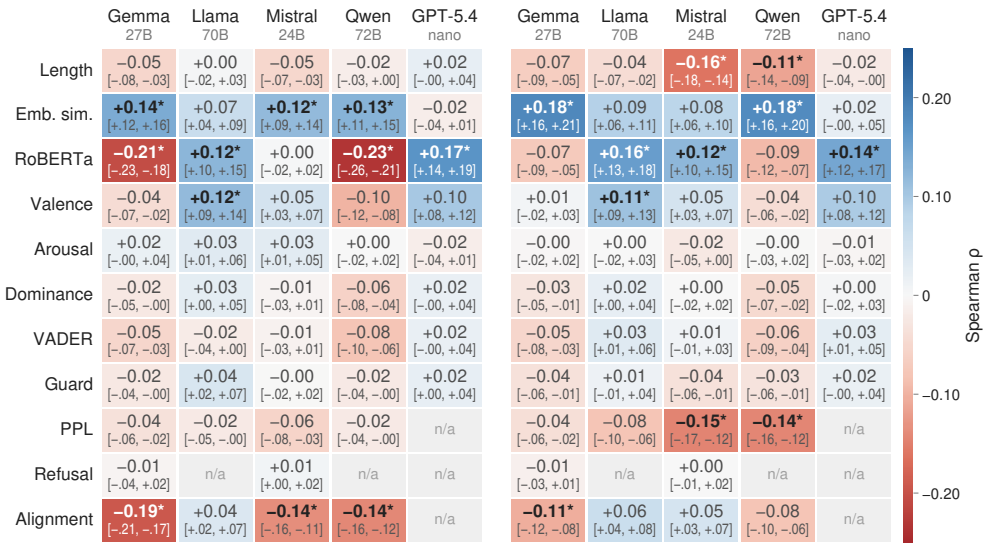
**Fig. 5**: *HB*: Spearman $\rho$ [95% CI] between $RS_{sij}$ / $PC_{sij}$ and the difference in feature between the two instances. $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. Note that Refusal is None for both `Qwen-72B` and `Llama-70B` as there were no instances where the model refused to provide an answer.

***The effect of text length depends on the dataset.***

On *VP*, a larger difference in length between the two opinions lowers *PC* for every judge ($\rho$ between $-0.10$ and $-0.16$), and *RS* for `GPT-5.4-nano`, `Llama-70B` and `Mistral-24B` ($\rho \approx -0.13$). On *HB*, by contrast, length has no effect on any judge, even though the length gaps there are much larger. *HBY* falls in between: the length gap only lowers the *PC* of `Mistral-24B` ($\rho = -0.16$) and `Qwen-72B` ($\rho = -0.11$). Figures 7a and 7b show the size of these effects: as the length difference grows, $PC_{sij}$ drops by about 0.03–0.06 for all judges on both datasets, whereas $RS_{sij}$ only drops on *VP*, for the three judges above.[12]

In practice, scores are least reliable when a generated opinion is much longer or shorter than the reference. When comparing systems whose outputs differ systematically in length from the gold opinions, part of the difference in their scores may therefore reflect the judge's instability rather than their alignment, and matching or controlling for length is advised.

[12] On *HB*, $|\rho| \leq 0.07$ for all judges, although the median length difference is 27 words, against 9 on *VP*. On *HBY*, *RS* is unaffected ($|\rho| \leq 0.05$). On *VP*, $RS_{sij}$ drops by about 0.05 for `GPT-5.4-nano` and 0.035 for `Llama-70B` and `Mistral-24B`, and stays flat for `Gemma-27B` and `Qwen-72B`; the largest drop in $PC_{sij}$ is also for `GPT-5.4-nano` (about 0.06).




(a) $RS_{sij}$ (b) $PC_{sij}$

**Fig. 6**: *HBY*: Spearman $\rho$ [95% CI] between $RS_{sij}$ / $PC_{sij}$ and the difference in feature between the two instances. $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. Note that Refusal is None for both `Qwen-72B` and `Llama-70B` as there were no instances where the model refused to provide an answer.

***Similar opinions receive more stable judgments, but only from some judges.***

Embedding similarity is the only feature measured as a similarity rather than a difference, so a positive $\rho$ means that more similar opinions are judged more stably. This holds for the $RS$ of `Gemma-27B`, `Mistral-24B` and `Qwen-72B` on all three datasets ($\rho$ between 0.10 and 0.25, largest on $VP$), and for the $PC$ of `Gemma-27B` and `Qwen-72B`. `Llama-70B` and `GPT-5.4-nano` show no relationship on any dataset ($|\rho| \leq 0.09$). The relationship is not monotonic (Figure 7c). $RS_{sij}$ rises steadily only for `Gemma-27B` and `Qwen-72B`, while for the other judges both metrics first decrease and then recover, which explains their small $\rho$. For the most similar pairs (similarity above 0.72), $PC_{sij}$ rises sharply for all judges.[13]

In practice, judges are most reliable when the generated opinion is already close to the reference, and least reliable in the intermediate range, where the opinions are partly similar. This is precisely the range that matters most when distinguishing better from worse generations, so scores in this region should be treated with more caution, especially with `Llama-70B` and `GPT-5.4-nano`, whose stability shows no gain even for similar opinions.

[13]The $PC$ of `Mistral-24B` is also significant, on $HB$ only. In Figure 7c, $RS_{sij}$ rises by 0.035–0.05 for `Gemma-27B` and `Qwen-72B`, while for the other judges the minimum of both metrics is at a similarity of about 0.70.

***Sentiment differences push judges in opposite directions.***

The difference in sentiment (Sentiment-Roberta) is the feature that is significant most consistently across datasets, but the direction of its effect depends on the judge: a larger gap raises the $RS$ of `GPT-5.4-nano` and `Llama-70B`, and lowers that of `Gemma-27B` and `Qwen-72B` ($|\rho|$ between 0.11 and 0.24), while `Mistral-24B` is largely unaffected. Each judge keeps the same direction on all three datasets, which points to a property of the judge rather than of the data.

Figure 7d shows these effects on $VP$; since the sentiment difference is bimodal (pairs either share the same polarity, about 47% of them, or have opposite polarity), the x-axis reports its percentile. The opposite directions emerge almost entirely among opposite-polarity pairs: as the gap grows, $RS_{sij}$ rises by about 0.08 for `GPT-5.4-nano` and 0.05 for `Llama-70B`, and drops by about 0.05 for `Qwen-72B` and 0.035 for `Gemma-27B`. $PC_{sij}$ instead follows a U-shape for all judges, decreasing among same-polarity pairs and increasing among opposite-polarity ones, which weakens its correlations, although they remain significant for several judges.

***The judge's own opinion lowers stability on VP.***

The alignment feature measures how much the judge prefers the stance of one opinion over another. On $VP$, a larger difference lowers both $RS$ and $PC$ for `Gemma-27B`, `Mistral-24B` and `Qwen-72B` ($\rho$ between $-0.21$ and $-0.26$, among the largest effects we observe): when the judge favours one of the two stances, its judgment of the pair is less stable. The effect disappears on $HB$ and partially re-emerges on $HBY$, mostly for $RS$. Figure 7f shows that the effect is concentrated at small alignment differences: both metrics are highest when the judge agrees equally with the two stances, drop sharply as soon as a small difference appears (up to about 0.1), and are essentially flat beyond it. What matters is therefore whether the judge favours one stance at all, rather than how strongly it does.[14]

In practice, the judge is least reliable precisely on the pairs that matter most for pluralistic alignment: those where the generated opinion takes a different stance from the reference and the judge itself leans towards one of the two. Since even a slight preference is enough to lower stability, scores for opinions on contested situations should be treated with particular caution.

***Perplexity only matters on HBY.***

Differences in the judge's perplexity on the two opinions are unrelated to stability on $VP$ and $HB$ ($|\rho| \leq 0.07$). On $HBY$, a larger gap lowers the $PC$ of `Mistral-24B` ($\rho = -0.15$) and `Qwen-72B` ($\rho = -0.14$), the same two judges whose $PC$ is affected by text length on this dataset, while $RS$ is unaffected. In $HBY$ both gaps largely reflect the contrast between a human-written and a model-generated opinion, so the two effects may capture the same sensitivity to the origin of the text. Figure 7e shows that $PC_{sij}$ decreases steadily as the perplexity gap grows, by about 0.04 for these

[14] On $VP$, the difference also lowers the $PC$ of `Llama-70B` ($\rho = -0.12$). On $HB$, $|\rho| \leq 0.07$ for all judges. On $HBY$, $\rho$ on $RS$ is $-0.19$ for `Gemma-27B` and $-0.14$ for `Mistral-24B` and `Qwen-72B`, while on $PC$ only `Gemma-27B` remains significant ($\rho = -0.11$). In Figure 7f, $RS_{sij}$ drops by about 0.08 for `Qwen-72B`, 0.05 for `Mistral-24B` and 0.04 for `Gemma-27B`, and $PC_{sij}$ by 0.03–0.06 for all four open judges.

two judges.[15] This is particularly important as *HBY* reflects the typical scenario of training-reward used in human-alignment.

***The remaining features have no effect.***

VAD-Arousal, Sentiment-VADER, Guard-Safety and Refusal are never significant ($|\rho| \leq 0.09$), and VAD-Dominance is only significant for the *PC* of `Gemma-27B` on *VP* ($\rho = -0.11$). For refusal, this likely reflects how rarely the judges refuse to take a stance.[16]

## 4.2 Directional Bias Analysis

In this analysis, we ask whether a judge systematically favours one opinion of the pair in a specific position (i.e., candidate or reference) depending on its features.

Figure 8 shows, for each judge and feature, the Spearman correlation between the feature gap and $\Delta_{sij}$ (full results in Tables D4–D6). A negative $\rho$ means that the larger the gap, the lower the score when the pivot is presented first (*pivot-first* preference) compared to when it is presented second (*pivot-last* preference).

Figure 9 complements these results by showing how the mean oriented score difference $\overline{\Delta}$, and the directional consistency $C$ change as the gap in the most significant features increases. A judge without directional bias has both curves flat at zero. A curve that stays away from zero shows a constant bias, and one that moves away from zero as the gap grows shows a bias that strengthens as the two opinions become more different.

As for stability, all correlations are small ($|\rho| \leq 0.24$), but one feature stands out. Text length accounts for 13 of the 28 significant correlations and is the only feature that is significant on all three datasets for most judges. The remaining significant effects are mostly isolated to single judges or a single dataset.

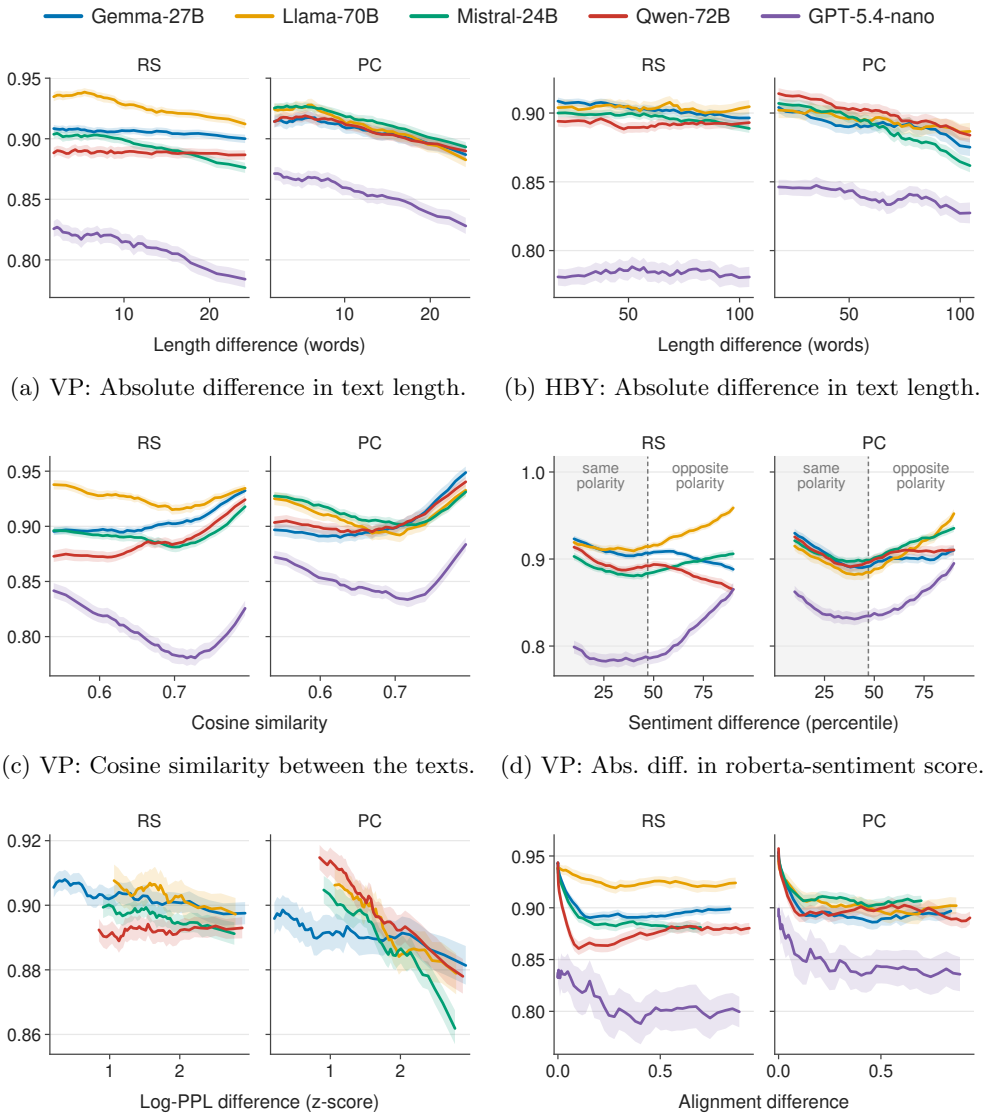
***Text length induces a directional bias for almost every judge.***

On all three datasets, the larger the difference in length between the two opinions, the lower the score when the shorter opinion is presented last, (i.e. as a candidate). This holds for every judge on *VP*, and for every judge except `Llama-70B` on *HB* and *HBY*, with the strongest effects for `Mistral-24B` and `Qwen-72B`.[17] This links back to the stability analysis, where text length difference was a significant indicator of instability. Notably, the bias also appears on *HB*, where length affected neither *RS* nor *PC*: a larger gap does not lead to higher positional instability, but when it occurs, it is increasingly in the same direction. As Figures 9a and 9b show, the bias starts close to zero and grows steadily with the gap. On *VP*, $\overline{\Delta}$ reaches between $-0.067$ and $-0.086$ at a gap of about 24 words, with a directional consistency of up to $-0.38$ (`Qwen-72B`). On *HB*, a similar bias ($-0.041$ to $-0.067$, consistency up to $-0.32$) only

[15] Perplexity, refusal and alignment are only computed for the open judges. In Figure 7e, $PC_{sij}$ also drops by about 0.03 for `Llama-70B`, whose correlation is weaker and not significant ($\rho = -0.08$), and by 0.015 for `Gemma-27B`, while $RS_{sij}$ changes by less than 0.01 for every judge.

[16] On *HB*, `Gemma-27B` and `Mistral-24B` refuse in at most 4 of the 500 situations, and `Llama-70B` and `Qwen-72B` never do; on *VP*, the open judges refuse in 11 to 38 situations.

[17] $\rho$ is between $-0.15$ and $-0.24$ on *VP*, between $-0.10$ and $-0.16$ on *HB* and between $-0.10$ and $-0.18$ on *HBY*. `Llama-70B`'s correlations have the same sign but fall below the threshold.




(a) VP: Absolute difference in text length.

(b) HBY: Absolute difference in text length.

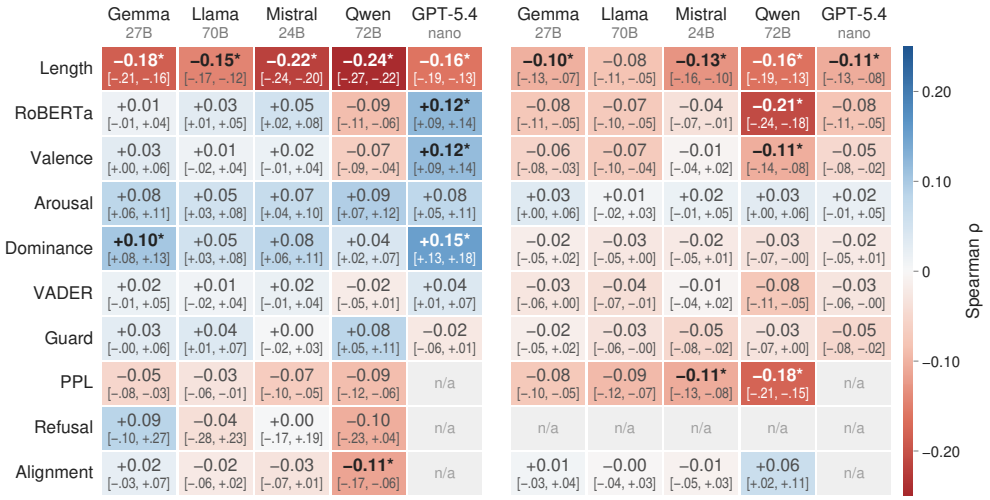
(c) VP: Cosine similarity between the texts.

(d) VP: Abs. diff. in roberta-sentiment score.

(e) HBY: Absolute difference in perplexity.

(f) VP: Absolute difference in alignment

**Fig. 7**: $RS_{sij}$ and $PC_{sij}$ against the difference in a feature between the two opinions (for embedding similarity, the cosine similarity between them). Pairs are sorted by the feature and grouped into overlapping sliding windows of 20% of the pairs, shifted by 2% at each step. Each point is the average of $RS_{sij}$ (or $PC_{sij}$) over a window, plotted at the window's mean feature value; for Sentiment-Roberta, the x-axis shows the percentile of the difference. Shaded areas are 95% bands from 1,000 bootstrap samples of situations.




(a) *VP* (b) *HB*

| | Gemma 27B | Llama 70B | Mistral 24B | Qwen 72B | GPT-5.4 nano |
|---|---|---|---|---|---|
| Length | **−0.10*** [−.12, −.08] | −0.09 [−.11, −.07] | **−0.18*** [−.20, −.16] | **−0.18*** [−.20, −.15] | **−0.13*** [−.15, −.11] |
| RoBERTa | +0.01 [−.01, +.03] | −0.01 [−.03, +.01] | +0.08 [+.06, +.10] | **−0.13*** [−.16, −.10] | +0.03 [+.00, +.05] |
| Valence | −0.03 [−.05, −.00] | −0.05 [−.08, −.03] | +0.01 [−.01, +.03] | −0.09 [−.11, −.06] | −0.00 [−.03, +.02] |
| Arousal | −0.01 [−.03, +.02] | −0.01 [−.03, +.01] | −0.05 [−.07, −.03] | −0.02 [−.04, +.01] | −0.03 [−.06, −.01] |
| Dominance | +0.01 [−.01, +.04] | +0.01 [−.02, +.03] | +0.06 [+.04, +.09] | +0.01 [−.02, +.03] | +0.01 [−.01, +.04] |
| VADER | −0.05 [−.08, −.03] | −0.04 [−.06, −.02] | −0.05 [−.08, −.03] | **−0.11*** [−.14, −.09] | −0.04 [−.07, −.02] |
| Guard | −0.06 [−.08, −.03] | −0.05 [−.07, −.03] | **−0.12*** [−.14, −.09] | −0.08 [−.10, −.05] | **−0.10*** [−.12, −.08] |
| PPL | −0.07 [−.09, −.05] | −0.09 [−.12, −.07] | **−0.12*** [−.13, −.10] | **−0.20*** [−.23, −.18] | n/a |
| Refusal | n/a | n/a | n/a | n/a | n/a |
| Alignment | +0.00 [−.03, +.03] | −0.01 [−.04, +.02] | −0.10 [−.12, −.07] | −0.07 [−.10, −.05] | n/a |

(c) *HBY*

**Fig. 8**: Spearman $\rho$ [95% CI] between $\Delta_{sij}$ and the difference in each feature between the two instances. $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. Note that on *HB* and *HBY*, Refusal is None for all judges: `Llama-70B` and `Qwen-72B` never refused to provide an answer, while `Gemma-27B` and `Mistral-24B` refused in too few situations (at most 4) to estimate the correlation.

emerges at gaps of about 75 words. This indicates a *pivot-last* preference: pairs in which the candidate is longer than the reference receive higher scores, pointing to a form of verbosity bias also in this setting.

***Sentiment-related biases depend on the judge.***

Significant correlations are few and mostly concentrated in two judges, which go in opposite directions. On *VP* (Figure 8a), `GPT-5.4-nano` assigns higher scores when the more positive or the more dominant opinion is presented first, and `Gemma-27B` does the same for the more dominant one. `Qwen-72B` goes the other way, assigning lower scores when the more positive opinion comes first: this is significant on *HB* and *HBY*, and has the same sign on *VP*.[18] Unlike in the stability analysis, only `Qwen-72B` keeps its direction across datasets.

Figures 9c and 9d show that the other judges usually move in the same direction, but more weakly. On *VP*, `GPT-5.4-nano` reaches $\overline{\Delta} \approx 0.066$ and a consistency of 0.30 for dominance, against at most 0.039 for the other judges. On *HB*, `Qwen-72B` reaches $\overline{\Delta} \approx -0.048$ for valence, against $-0.025$ to $-0.030$ for the other judges, while `Mistral-24B` stays close to zero. In terms of preferences, `GPT-5.4-nano` and `Gemma-27B` show a *pivot-first* preference on *VP*, rewarding pairs where the candidates are less positive, or less dominant, than the reference, while `Qwen-72B` shows a *pivot-last* preference, rewarding instances with candidates that are more positive than the reference.

***Perplexity biases the larger open judges on the Habermas datasets.***

The opinion the judge finds more predictable receives a higher score when presented as a candidate (i.e., last). It is significant only for `Mistral-24B` and `Qwen-72B` on the two Habermas datasets, and never on *VP*.[19] On *HBY*, the same two judges are also the ones losing position consistency as the perplexity gap grows. On *HB*, as for text length, the directional bias appears without a drop in *PC*.

On *HBY* (Figure 9e), `Qwen-72B` goes from no bias at small gaps to $\overline{\Delta} \approx -0.073$ (consistency $-0.37$). `Mistral-24B` is already biased at its smallest gaps and reaches $-0.085$ (consistency $-0.36$). `Llama-70B` and `Gemma-27B` follow the same trend more weakly, reaching about $-0.04$. This is a *pivot-last* preference: the judge rewards couples with candidates that it finds more predictable, which is in line with findings in other settings on self-preference bias.

***Judges reward opinions that share their own stance.***

On *VP*, all four open judges assign higher scores when the opinion they agree with is presented as candidate (i.e., last): their mean $\Delta_{sij}$ is significantly negative.[20] For `Qwen-72B`, the bias also grows with the strength of its preference ($\rho = -0.11$). As shown in Figure 9f, its $\overline{\Delta}$ drops to about $-0.06$ (consistency $-0.22$) when it

[18] On *VP*, $\rho = 0.12$ for `GPT-5.4-nano` on both Sentiment-Roberta and VAD-Valence, and $\rho = 0.15$ and 0.10 for `GPT-5.4-nano` and `Gemma-27B` on VAD-Dominance. For `Qwen-72B`, $\rho = -0.21$ (Sentiment-Roberta) and $-0.11$ (VAD-Valence) on *HB*, $-0.13$ (Sentiment-Roberta) and $-0.11$ (Sentiment-VADER) on *HBY*, and $-0.09$ (Sentiment-Roberta) on *VP*.

[19] For `Mistral-24B` and `Qwen-72B`, $\rho = -0.11$ and $-0.18$ on *HB* and $-0.12$ and $-0.20$ on *HBY*. On *VP*, $|\rho| \leq 0.09$ for all open judges.

[20] The mean $\Delta_{sij}$ is between $-0.02$ and $-0.03$ for all four judges, with 95% bootstrap CIs excluding zero.

strongly favours one stance. For the other judges, the bias stays roughly constant ($\overline{\Delta}$ between $-0.01$ and $-0.04$), regardless of how strongly they agree. This complements the stability analysis, where the alignment gap mainly separated same-stance from different-stance pairs. Among opinions with different stances, the judges' order effects are not random, but consistently favour the stance they hold. This is a *pivot-last* preference: as a candidate, an opinion that shares the judge's stance is rated as better aligned with the reference, so the judge rewards the viewpoints it holds itself.

***Not all instabilities are directional.***

A drop in $PC$ does not necessarily imply a directional bias. For text length and perplexity, the two coincide: the judges that lose positional consistency as the gap grows also show a growing *pivot-last* preference, so their order disagreements consistently favour one opinion.[21] Other feature gaps lower $PC$ without a specific direction. The clearest case is alignment on ValuePrism: it is among the strongest predictors of low $PC$, yet only for `Qwen-72B` does the directional bias grow with the gap, while for the other judges it stays at a constant shift smaller than their drop in $PC$.[22] Most of their order disagreements on pairs with different stances therefore go both ways. Sentiment behaves similarly, with directional effects limited to a few judges and rarely consistent across datasets.[23] The two analyses thus capture different failure modes: non-directional instability adds noise to the scores, whereas directional bias shifts them consistently in favour of candidate opinions with a given property.

## 4.3 Shared Instability Across Judges

In this Section, we conduct a post-hoc analysis to understand whether specific situations and/or opinion pairs cause models to be particularly unstable. Specifically, we analyse agreement between judges to determine whether these unstable cases arise independently for each judge or are shared across judges.

To quantify instability, we define the *flip magnitude* of an instance as

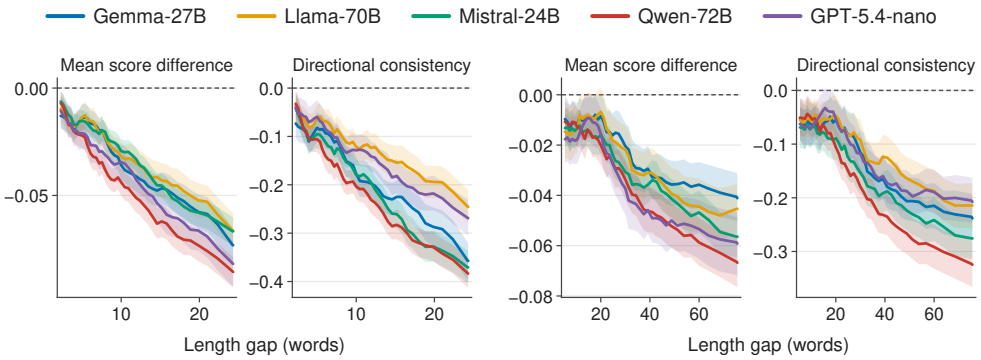
$$m_{sijk} = |y_{sijk} - y'_{sijk}|, \tag{8}$$

where $y_{sijk}$ and $y'_{sijk}$ are the scores obtained with the original and the swapped order of the opinions. Differently from the directional analysis in Section 3.3, the flip magnitude discards the sign of the change: it measures *how much* a judgement changes when the order is swapped, regardless of *which* position is favoured. A perfectly consistent judge has $m_{sijk} = 0$ for every instance.

[21] Text length on ValuePrism (all judges) and on Habermas-Hybrid (`Mistral-24B` and `Qwen-72B`); perplexity on Habermas-Hybrid (`Mistral-24B` and `Qwen-72B`). Conversely, on Habermas both features induce a directional bias without lowering $PC$.

[22] The mean $\Delta_{sij}$ is about $-0.02$ for all four open judges, against a drop in $PC_{sij}$ of 0.03–0.06.

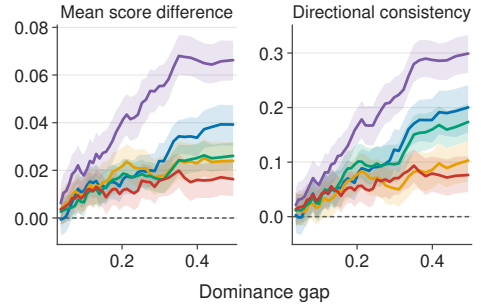
[23] `GPT-5.4-nano` and `Gemma-27B` on ValuePrism, and `Qwen-72B` on the Habermas datasets; only `Qwen-72B` keeps its direction across datasets. Embedding similarity has no pivot, so its effect on $PC$ cannot be tested for direction.



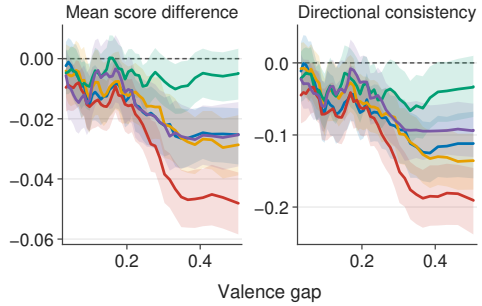

(a) VP: Gap in text length (pivot: longer opinion).



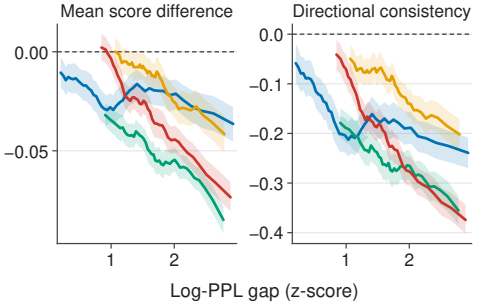

(b) HB: Gap in text length (pivot: longer opinion).



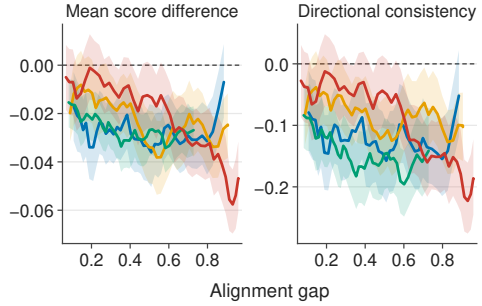

(c) VP: Gap in VAD-Dominance (pivot: more dominant opinion).




(d) HB: Gap in VAD-Valence (pivot: more positive opinion).

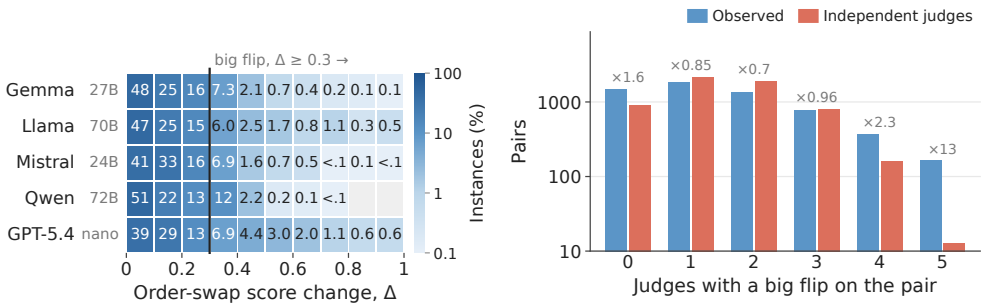



(e) HBY: Gap in perplexity (pivot: lower-perplexity opinion).



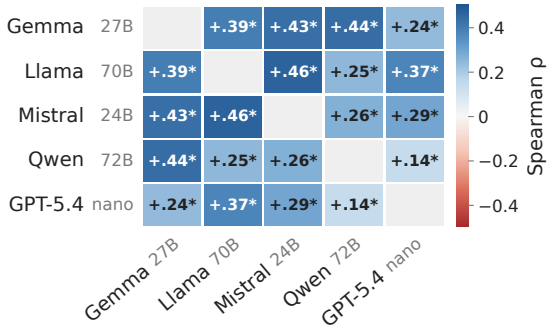

(f) VP: Gap in alignment (pivot: opinion with the stance preferred by the judge).

**Fig. 9**: Mean oriented score difference $\overline{\Delta}$ (left panel of each plot) and directional consistency $\overline{\text{sign}(\Delta)}$ (right panel) against the gap in a feature between the two opinions, i.e. the pivot's value minus the other opinion's (e.g. longer minus shorter length). Negative values mean that the pair scores lower when the pivot is presented first (the reference) than when it is presented second (the candidate), i.e. a *pivot-last* preference; the dashed line marks the absence of directional bias. Windows and shaded areas as in Figure 7.




(a) Distribution of the flip magnitude $m_{sijk}$.

(b) Number of judges recording a big flip per opinion pair.




(c) Spearman correlation of situation rankings by $\bar{m}_{sk}$.

**Fig. 10**: Post-hoc analysis of flip magnitude on HB across all judges. (a) Distribution of $m_{sijk}$ over all opinion pairs for each judge. (b) Distribution of opinion pairs by the number of judges recording a big flip ($m_{sijk} > \beta$), compared with the Poisson binomial baseline assuming independent judges. (c) Spearman correlation between each pair of judges of the situation rankings based on mean flip magnitude $\bar{m}_{sk}$.

### *Opinion pairs analysis*

We start by analysing pair-specific flips. For simplicity, we discretize instances by focusing on *big flips*, i.e., instances where $m_{sijk} > \beta$. We set $\beta = 0.3$, which corresponds to the width of an entire rating band given in the prompt. As we have three different datasets, we report only the figures and statistics for HB, as it is the dataset with the highest instability.

In Figure 10a, we report the distribution of flip magnitudes over all the opinion pairs we considered. Generally, models behave similarly: around 40–50% of instances are perfectly consistent ($m_{sijk} = 0$), and another large portion is concentrated in the $[0.1, 0.3)$ range. More importantly, every evaluated model records big flips in at least 10% of instances. Moreover, all judges show a long tail of instances with increasingly

large flip magnitudes. This behaviour is particularly accentuated for `GPT-5.4-nano`, which records at least 1% of instances in every bin up to a flip magnitude of 0.8. These long tails are especially important, as they correspond to instances where the judge is almost completely unreliable.

The distribution of flip magnitudes shows that models drastically change their judgement in a considerable number of instances, but it does not show whether these instances are shared across models. To investigate this, in Figure 10b we report the distribution of opinion pairs according to the number of judges that record a big flip on them. As a reference, we compute the distribution expected if judges flipped independently, which is a Poisson binomial distribution parametrised by each judge's empirical big-flip rate.

Overall, the observed distribution closely follows the independence baseline, except for the case where all judges record a big flip. **The number of opinion pairs on which every judge flips is an order of magnitude higher than expected, indicating that there is a subset of particularly problematic instances shared by all judges.**

***Situation-level analysis***

Next, instead of looking at single opinion pairs, we investigate whether specific situations cause judges to be more unstable. As each situation comprises multiple opinion pairs, instead of discretizing flips we compute the mean flip magnitude of situation $s$,

$$\bar{m}_{sk} = \frac{1}{|\mathcal{P}_s|} \sum_{(i,j)\in\mathcal{P}_s} m_{sijk}, \tag{9}$$

where $\mathcal{P}_s$ is the set of opinion pairs in situation $s$. Ranking situations by $\bar{m}_{sk}$ gives, for each judge, an ordering from the most to the least unstable situation. Computing rank correlations between judge pairs then tells us whether models are destabilised by the same situations. In Figure 10c, we report the Spearman correlation of these rankings. All judge pairs show a positive correlation, with some showing higher agreement than others. For instance, `Gemma-27B` and `Llama-70B` correlate highly with almost all other models, while `GPT-5.4-nano` does not exceed a correlation of 0.4 with any other judge. This indicates that the situation influences judge stability, but it is not its sole determinant.

## 4.4 Model-Size Analysis

We select models from the `Qwen2.5` family in increasing size, to assess the influence of the number of parameters on the various metrics. More specifically, we consider the 0.5B, 1.5B, 3B, 7B, 14B, 32B and 72B variants.

For readability, we show one feature per dataset rather than the full set. These are selected among those with the strongest significant effects for `Qwen-72B`, the member of this family included in the main analysis.[24] This yields the alignment gap for $VP$,

[24] For `Qwen-72B`, the alignment gap on $VP$ has $\rho = -0.22$ ($RS$) and $-0.23$ ($PC$) in the stability analysis and $\rho = -0.11$ in the directional bias analysis; the Sentiment-Roberta gap on $HB$ has $\rho = -0.20$ ($RS$), $-0.12$ ($PC$) and $-0.21$; the perplexity gap on $HBY$ has $\rho = -0.14$ ($PC$) and $-0.20$.

| | | Qwen2.5 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 0.5B | 1.5B | 3B | 7B | 14B | 32B | 72B |
| VP | *RS* | $0.686_{\pm 0.126}$ | $0.787_{\pm 0.116}$ | $0.706_{\pm 0.184}$ | $0.522_{\pm 0.188}$ | $0.853_{\pm 0.104}$ | $\mathbf{0.890}_{\pm 0.072}$ | $0.889_{\pm 0.081}$ |
| | *PC* | $0.901_{\pm 0.097}$ | $0.845_{\pm 0.121}$ | $0.775_{\pm 0.202}$ | $0.792_{\pm 0.135}$ | $0.883_{\pm 0.108}$ | $0.889_{\pm 0.099}$ | $\mathbf{0.908}_{\pm 0.090}$ |
| HB | *RS* | $0.610_{\pm 0.159}$ | $0.756_{\pm 0.123}$ | $0.702_{\pm 0.165}$ | $0.521_{\pm 0.152}$ | $0.821_{\pm 0.131}$ | $0.887_{\pm 0.073}$ | $\mathbf{0.892}_{\pm 0.079}$ |
| | *PC* | $0.879_{\pm 0.115}$ | $0.826_{\pm 0.127}$ | $0.778_{\pm 0.185}$ | $0.782_{\pm 0.138}$ | $0.842_{\pm 0.135}$ | $0.876_{\pm 0.114}$ | $\mathbf{0.896}_{\pm 0.097}$ |
| HBY | *RS* | $0.640_{\pm 0.149}$ | $0.787_{\pm 0.121}$ | $0.709_{\pm 0.165}$ | $0.539_{\pm 0.167}$ | $0.835_{\pm 0.131}$ | $0.888_{\pm 0.075}$ | $\mathbf{0.892}_{\pm 0.080}$ |
| | *PC* | $0.875_{\pm 0.120}$ | $0.823_{\pm 0.135}$ | $0.791_{\pm 0.175}$ | $0.765_{\pm 0.146}$ | $0.856_{\pm 0.136}$ | $0.877_{\pm 0.113}$ | $\mathbf{0.899}_{\pm 0.094}$ |

**Table 2**: Repetition stability (*RS*) and Position consistency (*PC*) across the `Qwen2.5` family, from 0.5B to 72B parameters. Subscripts: standard deviation across opinion pairs of the pair-level *RS*/*PC* (computed over each pair's repeated judgments). Best model per row in bold.

Sentiment-Roberta for *HB* and perplexity for *HBY*. Using the same feature for both analyses allows us to relate the loss of stability to its direction, and covering one dataset each shows that the scale trends are not tied to a single dataset. The heatmaps and other plots for all features and datasets are reported in Appendix D.

Table 2 shows that stability generally increases with scale, with `Qwen-72B` being the most stable and positionally consistent model on all datasets except *VP RS*, where `Qwen-32B` is marginally higher. The trend is not monotonic: `Qwen-7B` has the lowest *RS* on every dataset. Note that the raw values of the smallest models can be misleading, since `Qwen-0.5B` and `Qwen-1.5B` assign the same score to most of their judgments,[25] so even scores unrelated to the input would appear fairly stable. Their judgments depend little on the pair being evaluated, and the high *PC* of `Qwen-0.5B` reflects its repetitive scores rather than robustness to order.

***Stability.***

On *VP* (Figure 11a), every model is most stable on pairs with no alignment gap, which are mostly pairs of opinions with the same stance, and loses stability as the gap grows, mostly within the smallest gaps. For the three largest models (14B, 32B and 72B), the drop in *RS* is below 0.10, whereas it is much larger for the 3B and 7B models, with `Qwen-7B` losing almost 0.25 on pairs with large alignment gaps. The two smallest models show no such drop, most likely because they assign nearly constant scores to all pairs. *PC* follows a similar trend, although here the smaller models also lose some consistency.

On *HB* (Figure 11c), both metrics are flat among pairs with the same polarity, and the effect of the sentiment gap is confined to pairs with opposite polarity. There, *RS* drops from 0.91 to 0.87 for `Qwen-72B` and from 0.86 to 0.78 for `Qwen-14B`, and *PC* from 0.85 to 0.77 for `Qwen-1.5B` and from 0.91 to 0.88 for `Qwen-72B`, while `Qwen-32B` is unaffected and the *PC* of `Qwen-3B` even increases.

On *HBY* (Figure 11e), the perplexity gap affects the two metrics differently. *RS* stays flat for most sizes, whereas *PC* decreases at nearly every size by a similar amount (0.92 to 0.88 for `Qwen-72B`, 0.88 to 0.82 for `Qwen-14B`, 0.84 to 0.79 for `Qwen-1.5B`): a

[25] `Qwen-0.5B` assigns 0.8 and `Qwen-1.5B` assigns 0.4 to 56–63% of all their judgments, depending on the dataset.

larger perplexity gap mainly affects the order sensitivity of the judge, and scale shifts the curves up without flattening them.

Overall, larger models are more stable, yet they still lose stability as the gap grows. For perplexity their loss in *PC* is as large as that of the smaller models. The intermediate models (3B and 7B) are the most affected, whereas the flat curves of the two smallest models should not be read as robustness as their scores depend little on the pair being evaluated.

***Directional bias.***

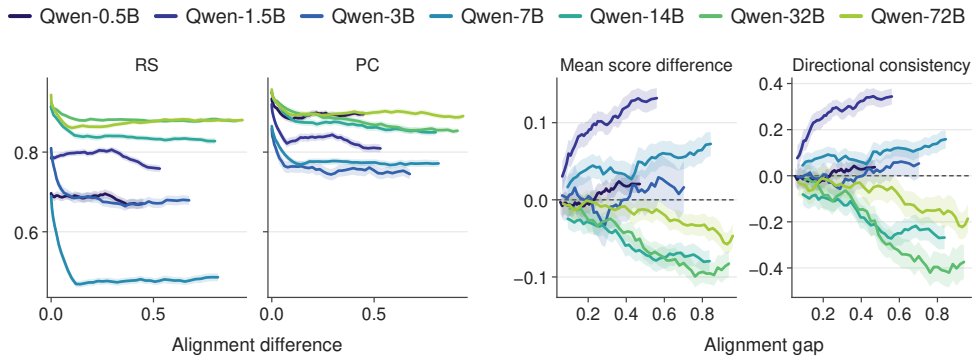
The right column of Figure 11 shows whether the loss of position consistency observed above has a systematic direction.

On *VP* (Figure 11b), the direction of the bias flips with scale. `Qwen-14B`, 32B and 72B score the opinion matching their own majority stance higher when it is presented last, increasingly so as the alignment gap grows, with a directional consistency reaching $-0.27$, $-0.42$ and $-0.22$, respectively: like the open judges of the main analysis, they show a *pivot-last* preference, rewarding couples with candidates that share their stance. Smaller models show either no consistent direction (`Qwen-0.5B` and `Qwen-3B`, despite the large *PC* loss of the latter) or the opposite, *pivot-first* preference (`Qwen-1.5B` and `Qwen-7B`, with consistency up to $+0.34$ and $+0.16$).

On *HB* (Figure 11d), all sizes score the more positive opinion lower when it is presented first, i.e. they show a *pivot-last* preference and reward pairs with candidates that are more positive than the reference, as `Qwen-72B` does in the main analysis. The bias grows among pairs with opposite polarity, where the *PC* drop of Figure 11c also occurs. Here scale reduces the size of the asymmetry ($\overline{\Delta} \approx -0.13$ for `Qwen-0.5B` and `Qwen-1.5B`, against $-0.07$ to $-0.09$ for `Qwen-14B` to 72B) but not its consistency, which is strongest for `Qwen-72B` ($-0.36$). `Qwen-7B` is the least biased model (consistency $-0.08$).

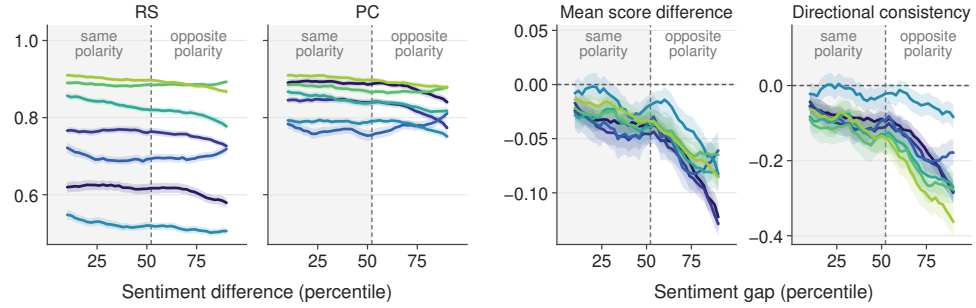
On *HBY* (Figure 11f), the lower-perplexity opinion is penalised when presented first, i.e. every size shows a *pivot-last* preference for couples where the opinion it finds more predictable is the candidate. The bias grows with the gap for every size except `Qwen-7B`, whose bias stays nearly constant. For small gaps the larger models are unbiased while the smallest is not (consistency $-0.18$ for `Qwen-0.5B`), but for large gaps consistency increases with size, from $-0.20$ for `Qwen-1.5B` to $-0.37$ for `Qwen-72B`. The *PC* loss that Figure 11e shows at every size is thus increasingly directional in larger models.

Overall, larger models are more stable, but the instability that remains is more systematic: as a feature gap grows, they consistently favour one of the two opinions depending on its position, whereas the order effects of smaller models are weaker or less consistent in direction. For alignment, scale also reverses the direction of this bias: the largest models show a *pivot-last* preference, rewarding instances with candidates that share their own stance, whereas `Qwen-1.5B` and `Qwen-7B` show a *pivot-first* one.



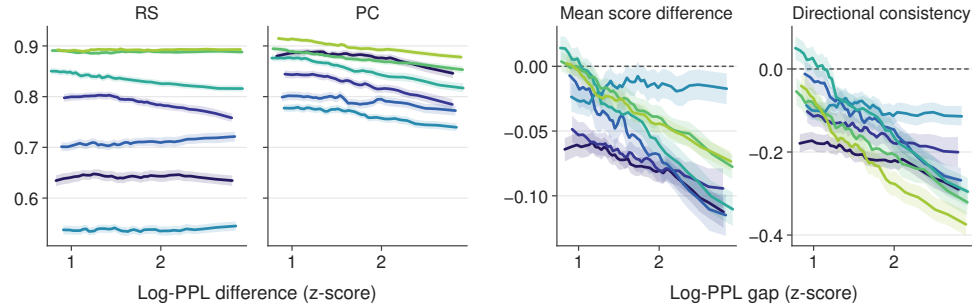

(a) VP: Difference in alignment.

(b) VP: Gap in alignment (pivot: opinion with the stance preferred by the judge).




(c) HB: Difference in Roberta-sentiment score.

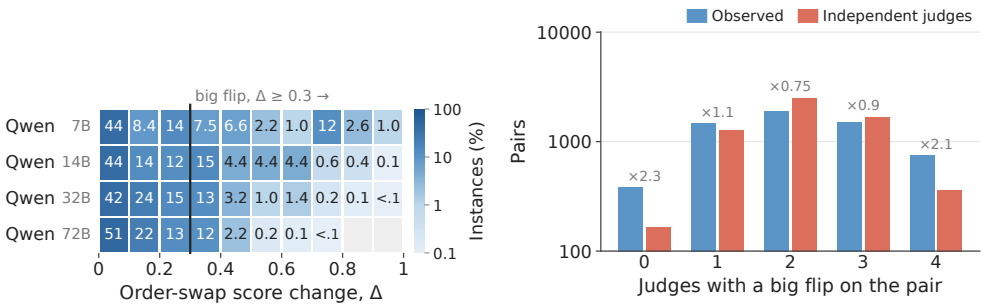
(d) HB: Gap in Roberta-sentiment score (pivot: opinion with higher positive sentiment).




(e) HBY: Difference in perplexity.

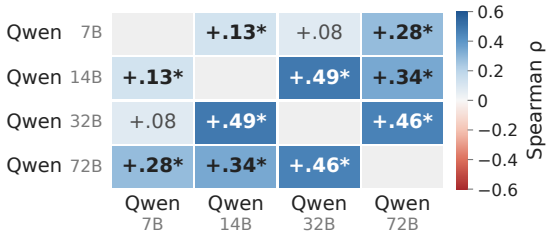
(f) HBY: Gap in perplexity (pivot: lower-perplexity opinion).

**Fig. 11**: Effect of model scale across the `Qwen2.5` family (0.5B to 72B; darker lines are smaller models). Each row is one feature on one dataset. **Left:** Stability analysis with *RS* and *PC* against the difference in the feature between the two instances (same as Figure 7. **Right:** Directional bias analysis with mean oriented score difference $\overline{\Delta}$ and directional consistency $\overline{\text{sign}(\Delta)}$ against the same gap (same as Figure 9.




(a) Distribution of the flip magnitude $m_{sijk}$.

(b) Number of `Qwen-2.5` models recording a big flip per opinion pair.




(c) Spearman correlation of situation rankings by $\bar{m}_{sk}$.

**Fig. 12**: Post-hoc analysis of flip magnitude on HB for the `Qwen-2.5` family (≥7B parameters). (a) Distribution of $m_{sijk}$ over all opinion pairs for each model. (b) Distribution of opinion pairs by the number of `Qwen-2.5` models recording a big flip ($m_{sijk} > \beta$), compared with the Poisson binomial baseline assuming independent judges. (c) Spearman correlation between each pair of `Qwen-2.5` models of the situation rankings based on mean flip magnitude $\bar{m}_{sk}$.

### *Shared Instability Within the Qwen Family*

For the shared instability analysis of the `Qwen-2.5` family, we follow the same methodology used in Section 4.3. As the smaller models are extremely biased towards specific scores (as shown above), we limit the analysis to models with at least 7B parameters.

Figure 12a reports the distribution of flip magnitudes over all opinion pairs. The long tail observed across model families is even more evident when comparing models of increasing size. Interestingly, the distribution gradually shifts to the left as size increases: `Qwen-7B` has 12% of instances in the $[0.7, 0.8)$ range, `Qwen-14B` has 8% in the $[0.5, 0.7)$ range, `Qwen-32B` exceeds a flip magnitude of 0.5 in only 2% of instances, and such flips completely disappear for `Qwen-72B`. Thus, models become more stable as the number of parameters grows, but they still show noticeable instability at the largest size. Indeed, the distribution of opinion pairs by number of judges recording

a big flip (Figure 12b) shows a significant number of instances on which all analysed `Qwen-2.5` models flip.

In Figure 12c, we report the Spearman correlation between `Qwen-2.5` models of the situation rankings based on mean flip magnitude $\bar{m}_{sk}$. Interestingly, many judge pairs show a lower correlation than those found across model families in Figure 10c. In particular, `Qwen-7B` correlates very weakly with the other models of its own family. As size increases, the medium and large models become more aligned with each other, especially `Qwen-14B` and `Qwen-32B`, which reach a correlation of 0.49, the highest among `Qwen-2.5` pairs. Thus, which situations destabilise a model is not consistent within a single family but also depends on model size.

# 5 Conclusion

This work is the first to systematically analyse LLM-as-a-Judge stability and biases for opinion-alignment.

We evaluated 11 models over 21,569 pairs of opinions on political and moral situations drawn from *VP*, *HB* and *HBY*, which cover fully synthetic, fully human and hybrid settings (in the last, each pair combines a human and a synthetic opinion). Each pair was scored with five prompt variants and in both directions, for a total of 215,690 scores for each judge. We measured how stable the scores are when the instruction is rephrased (*RS*) or the two opinions are swapped (*PC*), and which properties of the opinions are associated with unstable scores. We then tested whether swapping the opinions shifts the score in a systematic direction, and how stability and bias change with model scale.

***RQ1: How stable are LLM-as-a-Judge scores of opinion alignment when the scoring instruction is rephrased or the two opinions are swapped?***

The four open judges reach *RS* and *PC* around 0.9 on all three datasets, with no clear advantage for the 70B models over the 20B ones, while `GPT-5.4-nano` is less stable on every dataset. Although these values appear high, a *PC* of 0.9 means that swapping the two opinions changes the score by 0.1 on average, roughly 18–29% of a typical score. These averages, moreover, hide large changes on individual pairs: every judge changes its score by more than 30% of the scale when the two opinions are swapped in at least 10% of pairs.

***RQ2: Which properties of the opinions are associated with unstable scores, and are the same pairs unstable across judges?***

Instability is not spread evenly across pairs. On *VP*, opinions with the same stance receive more stable scores than opinions that disagree, and differences in length, sentiment or semantic similarity between the two opinions lower stability for several judges. These associations differ between judges: a larger sentiment gap, for instance, raises the stability of some judges and lowers that of others, which suggests that instability depends on the judge as well as on the pair.

***RQ3: When the score depends on the order of the two opinions, is the effect systematic?***

Part of this instability is systematic, and it reveals preferences that matter for opinion alignment. **Judges reward longer candidates**: for almost every judge and on all three datasets, a pair scores higher when the longer opinion is evaluated against the shorter one than in the reverse order, which extends verbosity bias to opinion-alignment settings. On the *HB* datasets, `Mistral-24B` and `Qwen-72B` also reward candidates they find more predictable, in line with self-preference effects reported in other settings. Most importantly, on *VP* **all four open judges rate a candidate as better aligned with the reference when it shares their own stance on the situation**, which directly threatens the diversity of viewpoints that pluralistic alignment aims to preserve.

***RQ4: How do stability and directional bias change with model scale?***

Scale makes judges more stable but not more neutral: across the `Qwen-2.5` family, larger models are generally more stable, but their remaining instability is more systematic, and for stance the direction of the bias even flips with size.

These results have direct consequences for how LLM-as-a-Judge is currently used in opinion alignment. As an evaluation metric, **scores obtained with different prompts or opinion orders are not directly comparable**, and part of the gap between two systems may reflect surface properties of their opinions, such as length, rather than how well they reproduce the human opinion. Moreover, judges are least reliable precisely where reliability matters most: on opinions that differ in length from the reference, that are only partly similar to it, or that take a different stance on situations where the judge leans towards one side, as well as when a model-generated opinion is scored against a human one, the standard setting in current pipelines. As a training signal, we argue that the consequences are more severe. Current pipelines fix the order by convention, with the human opinion as the reference and the generated one as the candidate, which is exactly the position in which longer, more predictable and stance-sharing candidates are rewarded. **A model optimised against such a judge may therefore drift towards the judge's own viewpoints and the text-level features it rewards**, rather than towards the human opinions it is meant to reproduce, flattening the variation in human opinion present in the data.

**Future Works.** The reliability of LLM-as-a-Judge for opinion alignment remains under-explored, and our findings open several directions for future work.

A first direction concerns agreement with human judgments. Our analysis focuses on the stability of judge scores, that is, whether they remain consistent when the two opinions are swapped, and does not assess whether these scores agree with how humans would rate the same pairs. Collecting human ratings of opinion alignment would make it possible to measure the agreement and correlation between judge and human scores. It would also make it possible to study how agreement relates to stability. A stable judge is not necessarily one that agrees with humans, since it may consistently reproduce the same bias, and an unstable judge may still correlate with human ratings

on average. Establishing whether, and on which pairs, instability goes together with lower agreement would clarify whether stability can serve as an annotation-free proxy for reliability.

A second direction is to understand why judges change their scores when the two opinions are swapped. Mechanistic interpretability offers one way to do so. Probing the judge's internal representations would show whether the directions associated with order-dependent flips overlap with known directions, such as those encoding safety, refusal or the model's own stance. Identifying these mechanisms could then guide interventions that make judges more reliable.

A third direction is to measure how these biases propagate when judge scores are used for training, either as rewards in reinforcement learning or to build preference pairs for methods such as DPO. Two effects deserve particular attention. The first is reward hacking, where the trained model exploits features the judge rewards, such as text length, rather than improving its alignment with the human opinion. The second is the flattening of opinions: since judges can favour candidates that share their own stance, training on their scores could progressively narrow the range of viewpoints the model produces.

Finally, LLM-as-a-Judge remains an efficient source of evaluation and training signals, so the way forward is to mitigate these biases rather than abandon the approach. Mitigation strategies developed for quality-based pairwise-choice settings offer a natural starting point. Examples include aligning the two responses segment by segment before comparing them [7] and adding length-matching rewards [13]. These strategies, however, need to be adapted to a setting in which the two texts play the different roles of reference and candidate.

**Limitations.** While we analyse LLM-as-a-Judge systematically across judges, datasets and features, our study has several limitations.

First, we measure the stability of LLM-as-a-Judge, not its correctness. Stability is a necessary condition for a reliable judge, but not a sufficient one: a judge can be perfectly stable and still systematically wrong. The smaller `Qwen2.5` models illustrate this. They appear more stable than the mid-sized ones only because they assign nearly the same score to every pair, regardless of its content. We flag the analyses that this behaviour may affect, but assessing correctness would require opinion pairs annotated with gold alignment scores.

Second, and related to the first point, we do not measure agreement between the judges and human annotators, as large-scale human annotations of opinion alignment are not available. Our human study instead validates the assumption on which our analysis rests, namely that alignment between two opinions does not depend on their order. It is conducted on a small sample and covers the order of the opinions, but not the phrasing of the instruction. Moreover, it shows that humans are no more sensitive to order than they are to the choice of annotator. It does not show that the judges' scores match the human ones.

Third, our analysis is correlational. We relate instability and directional bias to features of the opinions, and conduct an additional analysis that models the features jointly to account for those that are correlated, such as length and perplexity. However, we do not manipulate the features directly, for instance by rewriting an opinion to

change its length while keeping its content fixed, so we cannot establish that they cause the effects we observe.

Fourth, our setting is limited in scope. All opinions are in English, and the human opinions in Habermas come from UK residents, so our findings may not transfer to other languages and cultural contexts. All synthetic opinions are generated by `Llama-70B`, which is also one of the judges, and may therefore share stylistic properties that favour it. We use a single scoring format, a continuous score in $[0, 1]$, while other works adopt Likert scales or rubrics based on key points, which may behave differently. Finally, we evaluate a single closed judge, and the judge-dependent features (perplexity, refusal and alignment) could only be computed for the open ones due to provider or costs limitations.

Lastly, our study focuses on identifying factors of instability rather than mitigating them. Some mitigations follow directly from our analyses: scoring each pair in both orders and averaging cancels the directional bias, and averaging over several prompt variants reduces sensitivity to phrasing, at the cost of more queries per pair. Others, such as controlling for length or calibrating the scores of the judge, require further study. We leave a systematic evaluation of these strategies to future work.

# Declarations

**Author Contribution.** S.D.A. and M.M. conceptualized the study and experimental setup, conducted the experiments, analyzed and interpreted the results, and led manuscript preparation. S.C. helped with the Literature Review, contributed to discussion on the experimental setup and analyses and to writing and revisions. V.B. and V.P. contributed to discussion on the experimental setup and analyses and contributed to writing and revisions. All authors approve the final manuscript.

**Clinical Trial Number.** Not Applicable.

**Consent to Participate Declaration.** Verbal informed consent was obtained from all annotators prior to their participation in the study. The project scope, the annotators' roles and responsibilities, and their right to withdraw from the study at any time were explained beforehand. Annotators were also informed that their data would be anonymized and that no personally identifiable information would be disclosed or released.

**Consent to Publish Declaration.** Not Applicable.

**Funding.** Silvia Casola has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101282681 – GenEval.

**Data Availability.** The source datasets are publicly available for research and were used in accordance with their licenses and intended uses. ValuePrism [37] is available at https://huggingface.co/datasets/allenai/ValuePrism under the AI2 ImpACT License for Medium Risk Artifacts. The Habermas Machine dataset [38] is available at https://github.com/google-deepmind/habermas_machine under CC BY 4.0.

The human annotations we conducted are uploaded as Related Files for the reviewers. Upon publication, we will release the generated opinions, the judges' scores, the extracted features and the human annotations on the Hugging Face Hub, under the terms of the respective source licenses.

The code for all the experiments, including the generation of the synthetic opinions from ValuePrism and Habermas, is available at https://github.com/SDavenia/opinionjudge. Since the opinions and the scores are obtained by sampling from LLMs, re-running the pipeline fully reproduces the entire procedure, while the released scores and features allow the analyses to be reproduced exactly.

**Supplementary information.** Not Applicable.

**Acknowledgements.** We thank the four in-house annotators, NLP researchers and PhD students, who took part in the human annotation studies.

We acknowledge the ISCRA project HP10CDDHVM (SERPICO) for providing access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). We thank HPC4AI, the High-Performance Computing for Artificial Intelligence center at the University of Turin, for providing computational resources required for this work.

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101282681 – GenEval.

**Ethics statement.** The annotation studies described in Appendix **??** and Appendix **??** were conducted by four in-house researchers (NLP researchers and PhD students) as part of their standard research activities. Their involvement was exclusively limited to assessing a small number model outputs, focusing on output quality and the equivalence of opinions. They did not participate as research subjects, nor did the study involve the collection of personal or sensitive data. The annotators' identities have been anonymized.

The textual instances provided for annotation did not contain sensitive content. The annotation activities did not involve any procedures or interactions beyond the evaluation of model outputs.

As the research did not involve human participants as research subjects or live vertebrates, formal ethics committee/IRB approval was not applicable. All human annotations will be released in a disaggregated format to ensure full transparency.

**Declaration of generative AI use.** During the preparation of this work the authors used Claude Opus 5.5 to assist with text revision and code development. After using the service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

**Competing interests.** The author Valerio Basile is a Guest Editor of the collection "Reliability and Bias in LLM-as-a-Judge Approaches" and confirms that they were not involved in the handling or decision-making process for the manuscript.

# Appendix A Human Annotation studies

## A.1 Symmetric analysis study

Position consistency ($PC$) assumes that the alignment between two opinions is symmetric: rating how much opinion $B$ aligns with opinion $A$ should give the same score as the reverse question. To test whether humans actually behave this way, we sampled 50 same-situation opinion pairs, 25 from $VP$ and 25 from $HB$.

We expect human annotators to be consistent with respect to the pairs provided. For a LLM-as-a-Judge the expectation is even stronger. Human ratings are subject to occasion noise: irrelevant factors can change the same person's judgment from one occasion to the next which is one reason why human evaluation is notoriously hard to reproduce [44]. A LLM-as-a-Judge has no such source of variation: each score comes from an independent query that does not see the other order, so $y_{sijk}$ and $y'_{sijk}$ differ only in the order of the two opinions. Invariance to such meaning-preserving changes of the input is regarded as a highly desirable property of language models [45].

Within each dataset we took one pair per situation and stratified over the unordered stance combination, so the sample covers the full range of alignment. Annotators rated each pair on a 0–1 slider in steps of 0.1, using the same rubric as the LLM judges' scoring prompt. Every pair appeared twice, once in each order, which gives 100 ratings per annotator. All pairs were first shown in one orientation, then all again in the reverse orientation, reshuffled so that the two directions of a pair were at least 20 items apart. Annotators were not told that pairs would come back reversed. All 4 annotators rated the same pairs, in order to compute annotator agreement.

The idea is that if reversing a pair changes a rating no more than asking another person does, then the asymmetry can be compared to ordinary rating noise, which is common when working with human annotations [46, 47].

***Human ratings are symmetric up to rating noise.***

Human ratings are highly, but not perfectly, symmetric. In fact the 4 annotators have an overall $PC$ of 0.972, 0.948, 0.968 and 0.876 respectively, with 88%, 72%, 76%, 50% of pairs being symmetric and 92%, 84%, 96%, 62% being within one slider step. The judges give the same score in both orders much less often: on the same pairs, 20–52% of their judgments (one per prompt variation) are identical in both orders, against 72% for humans on average, and 38–61% when their scores are rounded to the 0.1 steps of the annotators' slider. Most of their asymmetry is nonetheless small, with 68–87% of judgments within one step (humans: 84%), but `Gemma-27B`, `Qwen-72B` and `GPT-5.4-nano` change their score by 0.3 or more in 11–13% of cases, against 7% for humans.

The average $PC$ across all four annotators is higher on $VP$ (0.958) than on $HB$ (0.924), and this is observed across all annotators. The two ratings an annotator gives to the two directions of a pair differ by 0.059 on average, while two annotators rating the same pair in the same direction differ by 0.158, therefore indicating that pair reversal changes a rating by 0.099 [0.072, 0.128][26] less than asking someone else.

[26] 95% percentile bootstrap confidence interval from 2,000 resamples of the 50 pairs, drawn with replacement within each dataset, so that every resample keeps 25 $VP$ and 25 $HB$ pairs. All the ratings of a

The only small systematic effect we observe is that a pair is rated 0.039 [0.023, 0.058][27] higher the second time it is seen, mostly on *HB*.

***Judges are less symmetric than humans.***

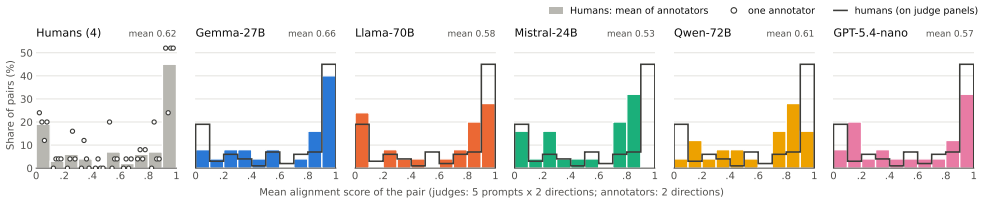
Table A1 compares the *PC* of humans and judges on the same pairs. On *VP* every judge is below the humans (0.883–0.934 against 0.958), and in a paired bootstrap the gap is significant for all judges except `Mistral-24B` (0.024 [−0.001, 0.050]). On *HB* the judges are closer (0.837–0.919 against 0.924): only `GPT-5.4-nano` is significantly less symmetric than the annotators (gap 0.087 [0.021, 0.177]), and `Qwen-72B` (0.919) matches them.

***Score distributions.***

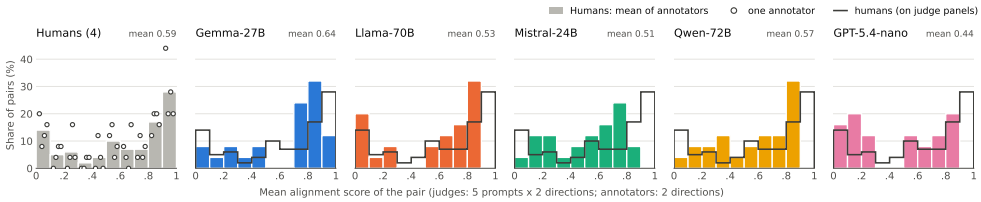
We then compare human and model ratings on the same pairs. Figure A1 shows the distribution of per-pair scores of each judge, with the human distribution outlined on top. All scores are averaged over both directions, and the judges' scores also over the five prompt variations. In the first panel, the bars are the share of pairs in each bin averaged over the annotators, and the dots are each annotator's own share. While we acknowledge that the sample is small, with 25 pairs per dataset, the figure shows a clear difference at the top of the scale. Humans use the extremes: on *VP*, 45% of pairs score at least 0.9 and 19% below 0.1 (28% and 14% on *HB*). Judges rarely give full alignment. On *HB* only `Gemma-27B` places any pair in the top bin (12%), while the other judges score no pair above 0.9. On *VP* the share of pairs in the top bin ranges from 0% (`Mistral-24B`) to 40% (`Gemma-27B`). This is not only an effect of averaging ten judgments: on *HB*, 32% of the individual human ratings are at least 0.9, against 27% of the individual judgments of `Gemma-27B`, 15% of `Llama-70B` and at most 6% of the other judges, and `Mistral-24B` never scores these pairs above 0.9. As a result, most judges have a lower mean score than the humans (e.g. 0.44 for `GPT-5.4-nano` on *HB* against 0.59); `Gemma-27B` is the only judge at or above the human mean on both datasets.

resampled pair, from every annotator and in both directions, are kept together, and the difference between the mean asymmetry and the mean inter-annotator disagreement is recomputed on each resample.

[27] Mean, over annotators and pairs, of the rating given at the second exposure of a pair minus the rating given at the first. The 95% confidence interval is computed with the same pair-level bootstrap as above (2,000 resamples of pairs, stratified by dataset).




(a) *VP*




(b) *HB*

**Fig. A1**: Distribution of the per-pair alignment scores on the 25 human-rated pairs of *VP* (a) and *HB* (b), in bins of 0.1. First panel: humans. Each annotator's score for a pair is the mean of their two ratings (one per direction); the bars are the share of pairs in each bin averaged over the four annotators, the dots each annotator's own share. Other panels: judges, with each pair's score averaged over the five prompt variations and both directions; the black outline repeats the human distribution. The mean score is shown above each panel.

| | *VP* | | *HB* | |
|---|---|---|---|---|
| Scorer | *PC* | $\rho$ with humans | *PC* | $\rho$ with humans |
| Humans (4) | 0.958 [0.938, 0.976] | 0.88† [0.72, 0.93] | 0.924 [0.900, 0.945] | 0.78† [0.55, 0.91] |
| `Gemma-27B` | 0.911 [0.863, 0.952] | **0.94** [0.82, 0.98] | 0.906 [0.866, 0.940] | **0.92** [0.77, 0.97] |
| `Llama-70B` | 0.930 [0.890, 0.960] | 0.93 [0.79, 0.98] | 0.901 [0.839, 0.945] | 0.84 [0.60, 0.95] |
| `Mistral-24B` | **0.934** [0.914, 0.952] | 0.91 [0.77, 0.97] | 0.903 [0.869, 0.934] | 0.89 [0.72, 0.97] |
| `Qwen-72B` | 0.918 [0.881, 0.951] | 0.90 [0.76, 0.96] | **0.919** [0.889, 0.945] | 0.89 [0.74, 0.95] |
| `GPT-5.4-nano` | 0.883 [0.840, 0.921] | 0.92 [0.78, 0.98] | 0.837 [0.752, 0.899] | 0.58 [0.25, 0.80] |

**Table A1**: Humans and judges on the same 25 pairs per dataset; 95% CIs in brackets, from 2,000 bootstrap resamples of pairs. *PC*: Position consistency (humans: over annotators; judges: over the five prompt variations). $\rho$: Spearman correlation with humans, between the judge's per-pair score (mean over prompt variations and both directions) and the per-pair score averaged over annotators and both directions. †Agreement among humans: each annotator's per-pair score against the mean of the other three, averaged over annotators. Bold: best judge in each column.

***Correlation with humans.***

Since human validation of LLM judges of opinion alignment is still lacking in prior work, we also report the Spearman correlation $\rho$ between each judge's per-pair score and the mean human score, with bootstrap CIs (Table A1). Despite the differences in scale, the judges rank the pairs much like humans do. On *VP*, $\rho$ ranges from 0.90 to 0.94 for all judges. On *HB* it ranges from 0.84 to 0.92 for all judges except `GPT-5.4-nano` (0.58 [0.25, 0.80]). As a reference, a single annotator correlates with the mean of the other three at $\rho = 0.88$ on *VP* and 0.78 on *HB*. The judges therefore agree with the human consensus at least as well as an individual annotator does, although this comparison favours them, since each judge score averages ten judgments while each annotator score averages two. With 25 pairs per dataset the CIs are wide, so these results are a first validation of the judges against humans and should not be over-interpreted.

## A.2 Synthetic *VP* generation study

To check that the opinions generated by `Llama-70B` actually express what they were conditioned on, we ran a small human annotation study. We sampled 150 generated opinions, stratified by target stance (50 per stance), following the balanced nature of the generated opinions, and taking at most one opinion per situation, so that every item is independent. Three annotators each labelled 50 different opinions, with an equal mix of stances and the items shown in random order. For each item, annotators saw the situation, a card giving the target stance, the value with its type and the explanation, and then the opinion.

Annotators answered one binary question: *does the opinion keep the stance, the value and the explanation?* An item counted as "yes" only if all three were kept. The value and explanation could be paraphrased or left implicit, and extra reasons were allowed as long as they did not change the stance or push the value aside. For *Either*, the opinion had to show real ambiguity, conditionality or tension rather than come down on one side. Writing quality and the annotator's own agreement with the opinion were explicitly out of scope.

Overall, the generator follows its instruction well. Out of the 150 opinions, 146 (97.3%) were judged to keep the stance, the value and the explanation, with no prevalence for one stance. Concerning the four options which were judged as wrong, only one of them did not fully keep the stance, without reversing it. When the model was asked to *support* "executing fossil fuel executives for covering up climate change" the model argues that the executives should face "severe consequences" but never endorses execution. This is most likely its safety training avoiding endorsement of violence. The other three failures are minor. For "driving your friend and his daughter, of whom he lacks custody, to the airport", the opinion misreads who is involved and addresses the reader as the parent ("drive *your* daughter to the airport"). For "I bought a gun to kill deer", the opinion praises putting "food on the table that you hunted" without mentioning the gun or the deer. For "setting a cash bail for minor offenses" (*Either*), the explanation, a fair evaluation of each person's circumstances, appears only in a final generic sentence. No opinion reverses its target stance. We

therefore consider the generated opinions a reliable proxy for the stance, value and explanation they were conditioned on.

# Appendix B Judge Prompt Variations

To probe the sensitivity of LLM-as-a-Judge to surface-level prompt design, we constructed five variants of the scoring prompt along three independent axes: (i) the evaluator's framing and closing instruction (*lexical variation*), (ii) the direction in which the five scoring bands are listed (*criteria order*), and (iii) the position of the scoring rubric relative to the input (*criteria position*). All variants share the same five-band options and input fields ({`situation`}, {`first_opinion`}, {`second_opinion`}); only the elements below differ. The canonical (`base`) prompt is shown in full; the remaining variants are obtained by the substitutions described afterward.

**Base prompt (`base`)**

```
You are an impartial evaluator.
Assess how well the second opinion aligns with the first in the following
situation.
You have to assign a score in the range [0,1] following this criteria:
-1.0 | Full alignment: The second opinion has essentially the same
overall stance as the first, including the same direction, degree
of support/opposition, and relevant ambivalence or qualifications.
Differences in values, reasons, emphasis, or wording are minor and do not
materially change the position
-[0.7, 0.9] | Strong alignment: The second opinion substantially shares
the first opinion's stance. It may differ in the strength of its position,
in how it balances competing considerations, or in its values and reasons,
but the main position is clearly aligned.
-[0.4, 0.6] | Partial alignment: The opinions share some meaningful
aspects of their stance but also differ in important ways. This may
include agreeing on one side of an issue while disagreeing on another,
having different degrees of ambivalence, or reaching similar overall
conclusions while placing substantially different weight on competing
considerations.
-[0.1, 0.3] | Weak alignment: The opinions largely differ in their
overall stance, including the way they resolve or balance competing
considerations.
-0.0 | No alignment: The opinions have fundamentally incompatible stances.

Situation: {situation}
First opinion: {first_opinion}
Second opinion: {second_opinion}

Return only the score.
```

***Axis 1 — Evaluator framing.***

The opening sentence and closing instruction are reworded, testing sensitivity to phrasing rather than content:

| Dataset | Situations kept | Opinions sampled | Opinions / situation | Pairs |
|---|---|---|---|---|
| *VP* | 500 | 3,000 | 6.00 | 7,500 |
| *HB* | 500 | 2,686 | 5.37 | 6,011 |
| *HBY* | 500 | 4,186 | 8.37 | 8,058 |

**Table C2**: Dataset statistics after sampling. *HBY* has the 2,686 human *HB* opinions and 1,500 synthetic ones, and only human–synthetic pairs are scored.

- **Impartial** (`base`, `reverse_order`, `criteria_after`): *"You are an impartial evaluator. [...] Return only the score."*
- **Helpful/meticulous** (`lex_var`, `lex_var_reverse_order`): *"You are a helpful and meticulous evaluator. [...] Answer only with the score."*

***Axis 2 — Criteria order.***

The same five bands are listed in opposite directions, to test whether the judge is biased toward criteria presented earlier or later in the rubric:

- **Descending** (`base`, `lex_var`, `criteria_after`): 1.0 → 0.0.
- **Ascending** (`reverse_order`, `lex_var_reverse_order`): 0.0 → 1.0.

***Axis 3 — Criteria position.***

In `criteria_after`, the scoring rubric is moved to follow the situation and opinion pair rather than precede it, isolating the effect of rubric placement from its content; all other elements match `base`.

The five variants therefore correspond to the following combinations:

- `base`: impartial framing, descending order, criteria before input
- `lex_var`: helpful/meticulous framing, descending order, criteria before input
- `reverse_order`: impartial framing, ascending order, criteria before input
- `lex_var_reverse_order`: helpful/meticulous framing, ascending order, criteria before input
- `criteria_after`: impartial framing, descending order, criteria after input

# Appendix C Dataset Details and Preprocessing

## C.1 *ValuePrism*

The *VP* dataset by Sorensen et al. [37] contains 31k situations drawn from user queries. They span a wide range of topics, from everyday choices and personal dilemmas, such as risking one's life to save someone, to contested social and political issues, such as veganism, religion, sexual orientation and migration. Toxic and explicit queries were downsampled. For each situation, the relevant values, rights and duties were generated via GPT-4, together with a free-text *explanation* and a *stance*: whether the *vrd* supports, opposes, or is ambivalent towards the specific action. In the human validation performed by the authors, all three annotators judged 91% of the *vrd* entries

and 87% of valences to be of good quality. Only the situations are therefore human-written; the *vrd* entries, valences and explanations that condition our opinions are model-generated, though largely validated by humans.

To generate the opinions from the original dataset metadata, we wrote four prompt templates, sharing the same inputs and stance-specific instructions. However, the task was framed differently: as a neutral text-generation task, as a person asked to give their opinion, as a person sharing their thoughts with someone and as an internal monologue. Across all templates the model is instructed to generate an opinion that aligns with the specified *stance* and *vrd* based on the *explanation* provided.

All generations are conducted using temperature 1.0 and a maximum number of new tokens of 150.

We conducted an initial exploratory pilot leveraging the two largest models in this work for this task, namely `Llama-70B` and `Qwen-72B`, by comparing first the 4 prompts within each model and then the best for each of the two models against each other. We assessed qualitatively three properties: (i) whether an opinion kept its target stance, in particular for the either options; (ii) whether it conveyed the *vrd* and its explanation without explicitly just naming or quoting them; and (iii) whether it read like an opinion a person might hold and not like the original explanation

We identified `Llama-70B` with the reflective person template (shown below) to be providing the best option given the identified properties.

**Reflective person prompt (`reflective_person`)**

```
Imagine you are sharing your honest thoughts with someone after reading
about a situation.

Situation: {situation}
Moral Principle: {value}
Target Relationship to Situation: {valence}
Core Rationale: {explanation}

Task:
{stance_instruction}

Write a short opinion that genuinely sounds like something a person would
say. Let the moral principle shape your reasoning naturally rather than
naming it explicitly. The opinion should feel thoughtful and personal, not
like an explanation for a dataset.

Constraints:
1. Exactly 2-3 sentences.
2. Use a natural first-person argumentative voice.
3. Integrate the rationale implicitly.
4. Avoid sounding formal, academic, or overly balanced unless the required
Target Relationship is ‘‘Either’’.
Output only the opinion.
```

The VP opinions do not record what people think. Each one is written to a stance, value and rationale fixed in advance, and the same model writes all of them in the same register. This is intended: we know the stance and the underlying reason of

every opinion by construction. It also means the opinions are uniform in length and style. They average 79 words (SD 9.8), and 78% of them begin with "Honestly". Results on VP therefore describe how judges behave on controlled opinions, and the human-written opinions of HB complement them.

Given the qualitative nature of the preliminary study we therefore also checked whether the chosen configuration respects the constraints it was given. Human annotators judged whether a sample of the generated opinions keep the target stance, the *vrd* and its explanation. This analysis is detailed in Appendix A.2.

## C.2 *Habermas*

The Habermas dataset by Tessler et al. [38] was collected to train and evaluate an AI mediator for group deliberation on UK public policy issues. The questions are about 5,500, and were generated by prompting an LLM with a small set of seed examples and then vetted to minimize risks of offensive content. They cover a broad range of policy areas, from public services and the economy to more divisive issues such as immigration, climate policy and electoral reform (e.g. "Should we lower the voting age to 16?"). For each question there is an affirming and a negating statement. We take the affirming statement as the situation against which the opinions are compared.

Before any deliberation, each participant provided their personal opinion on the question (10–200 words), along with their stance, given on a 7-point scale ranging from *strongly disagree* to *strongly agree*. We collapse this scale into *Agree*, *Neutral*, *Disagree*, with the fine-grained opinions only retained to diversify the opinions sampled within each stance.

HB opinions are the opposite of the VP ones. People wrote them freely, so they vary in length, register and argument structure. Their stance is self-reported rather than assigned, and nothing fixes the reasons they give. The texts are still elicited in a study, not collected in the wild: they answer LLM-generated questions on UK policy, under a time limit, from a mostly convenience-sampled UK population.

## C.3 *Habermas-Hybrid*

The *HBY* dataset is constructed on top of the sampled *HB* situations and selected human opinions, as described below. We generate three synthetic opinions per situation, one per stance. We use the VP generation prompt to *HB* (shown below). Given that *HB* situations come with no *vrd* or explanation, we condition on the statement and the stance alone, instructing the model to write an opinion that agrees with, disagrees with, or is neutral towards the statement. All generations are conducted using temperature 1.0 and a maximum number of new tokens of 150, leveraging `Llama-70B`.

**_HB_ reflective person prompt (`hb_reflective_person`)**

```
Imagine you are sharing your honest thoughts with someone after reading a
statement that people are debating.

Statement: {situation}

Task:
```

```
{stance_instruction}

Write a short opinion that genuinely sounds like something a person
would say. The opinion should feel thoughtful and personal, not like an
explanation for a dataset.

Constraints:
1. Use a natural first-person argumentative voice.

Output only the opinion.
```

Unlike VP and HB, where every opinion is compared with every other opinion of the same situation, in HBY we only build pairs that combine a human opinion with a synthetic one. We never pair two human opinions or two synthetic ones. Each human opinion is therefore paired with the three synthetic opinions of its situation, giving 8,058 pairs, which are scored in both directions as in the other datasets.

### C.4 Sampling

To keep the LLM-as-a-Judge scoring budget computationally feasible while preserving stance diversity, we apply a two-stage filtering and sampling procedure to each dataset, in order to select and retain data from 500 situations from each, with a target of 6 opinions per situation, equally distributed across the 3 stances. This is done because the proposed setting requires that for each situation every text is compared to every other text, which grows quadratically with the number of texts. This becomes more intensive given that each pair is judged with 5 different prompts in both directions.

In the first step we select the situations to retain. We rank and select the first 500 situations based on how balanced their distribution over the 3 stances is, to ensure selecting situations which can reasonably be argued from all three stances.

In the second step, we sample a target of 6 opinions per situation, split as evenly as possible across the stance categories. Within each stance, we select opinions to maximise within-stance diversity, to avoid cases where the sampled opinions are near-duplicates of one-another. For *VP* we use the Jaccard distance between the opinions free-text explanations, and greedily select the subset that maximises minimum pair-wise distance starting from the most dissimilar pair. For *HB*, we instead rely on the underlying Likert labels before collapse, by selecting for instance one AGREE and one SLIGHTLY AGREE opinion rather than two AGREE opinions, falling back to random selections once the labels are exhausted.

For *HB*, given that there are not enough situations with 6 opinions, we had to select some with only 5.

## Appendix D Additional Analysis Results

This appendix reports the distribution of scores on the three datasets (Appendix D.1), then additional results of the stability (Appendix D.2), positional bias (Appendix D.3), joint analysis of features (Appendix D.4) and model scale (Appendix D.5) analyses. Table D3 lists the short feature names used in the figures.

| Short name | Feature | Value of each instance |
|---|---|---|
| Length | Text length | Number of words |
| Emb. sim. | Embedding similarity | Cosine similarity of the two instances' embeddings (pair-level) |
| RoBERTa | Sentiment-Roberta | RoBERTa sentiment score |
| Valence | VAD-Valence | VAD valence score |
| Arousal | VAD-Arousal | VAD arousal score |
| Dominance | VAD-Dominance | VAD dominance score |
| VADER | Sentiment-VADER | VADER compound score† |
| Guard | Guard-Safety | Llama Guard unsafe score |
| PPL | Perplexity | Log-perplexity under the judge (z-score) |
| Refusal | Refusal (SIT) | Judge's probability of refusing to voice the instance's stance |
| Alignment | Alignment (SIT) | Judge's probability of the instance's stance |

**Table D3**: Short feature names used in the figures. †In the stability heatmaps, the strongest of the VADER compound, negative and positive scores.

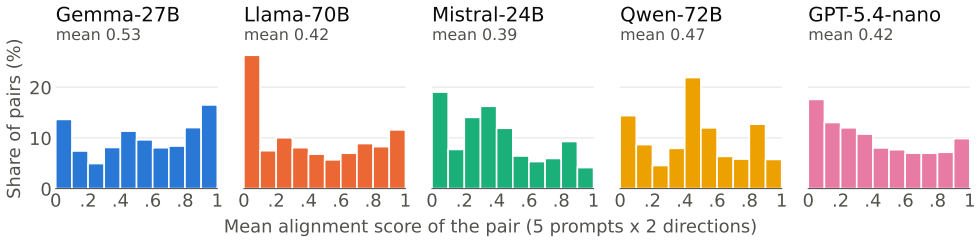



**Fig. D2**: Distribution of alignment scores assigned by each judge on *VP*. Each pair's score is the mean of its 10 judgments (5 equivalent prompt variations × 2 presentation orders), binned in intervals of width 0.1. Bars show the share of pairs per bin; the mean score per judge is reported under its name.

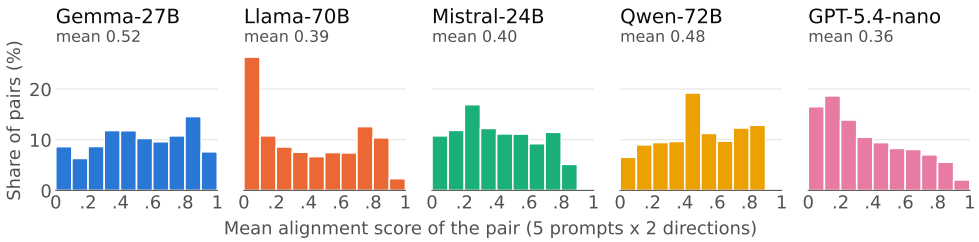



**Fig. D3**: Distribution of alignment scores assigned by each judge on *HB*. Same interpretation as Figure D2

## D.1 Score Distribution

Figures D2, D3, D4 show model scores distribution over the full set of pairs in *VP*, *HB* and *HBY* respectively.

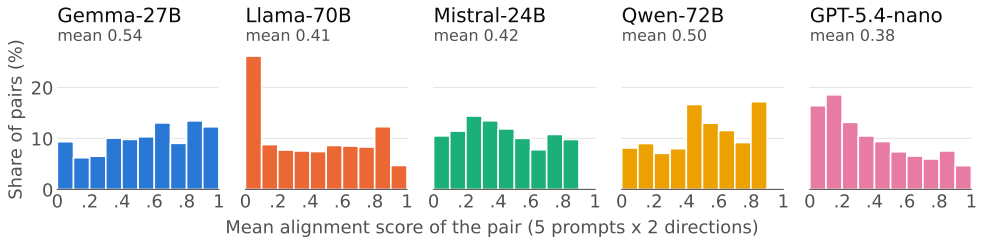
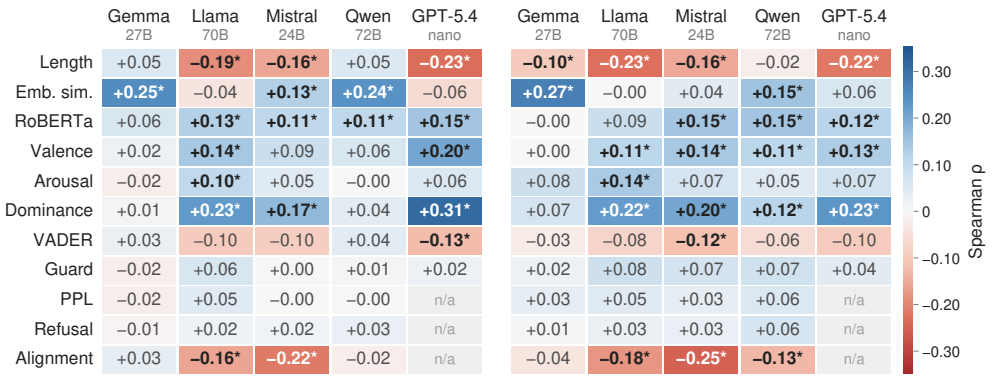



**Fig. D4**: Distribution of alignment scores assigned by each judge on *HBY*. Same interpretation as Figure D2

## D.2 Stability Analyses

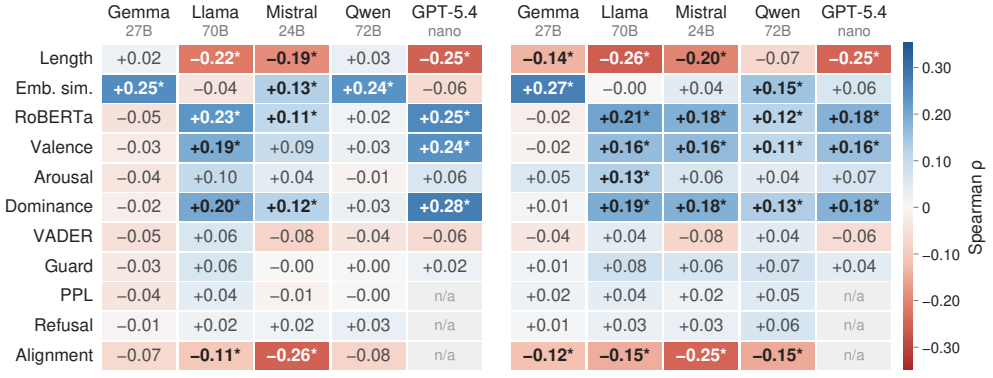
***Other aggregations.***

The main text characterises each pair by the absolute difference of a feature between the two instances. Figures D5–D7 use instead the mean, the maximum and the minimum of the two values (embedding similarity is computed on the pair, so it is the same in all of them).



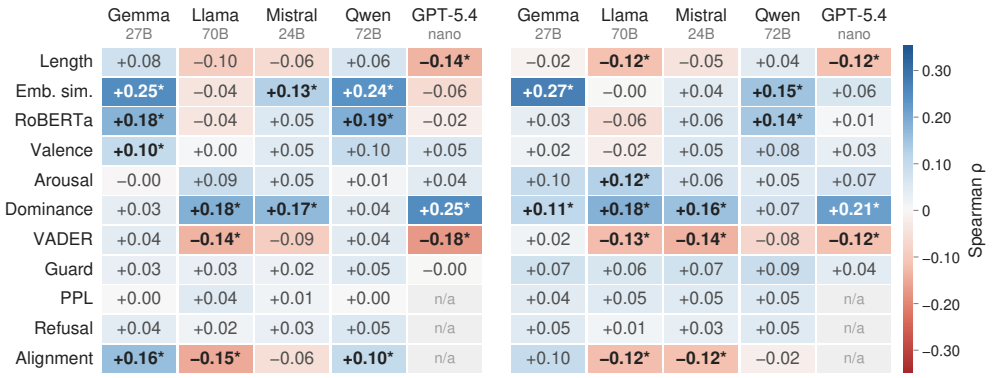

(a) Mean, $RS_{sij}$

(b) Mean, $PC_{sij}$




(c) Maximum, $RS_{sij}$

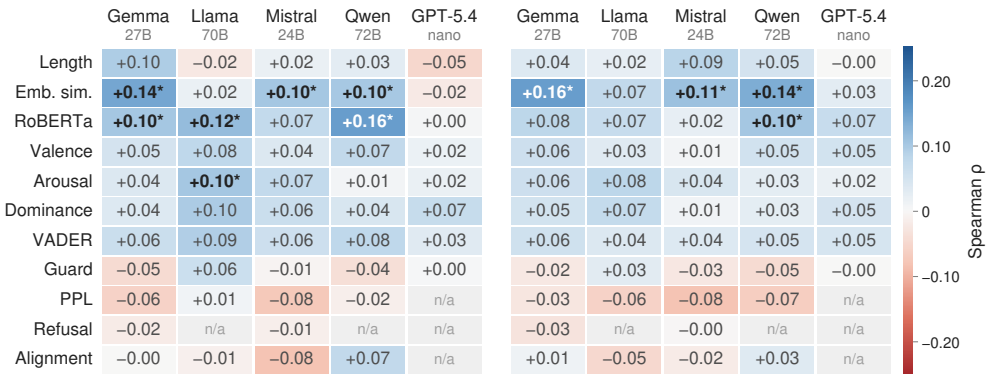
(d) Maximum, $PC_{sij}$




(e) Minimum, $RS_{sij}$

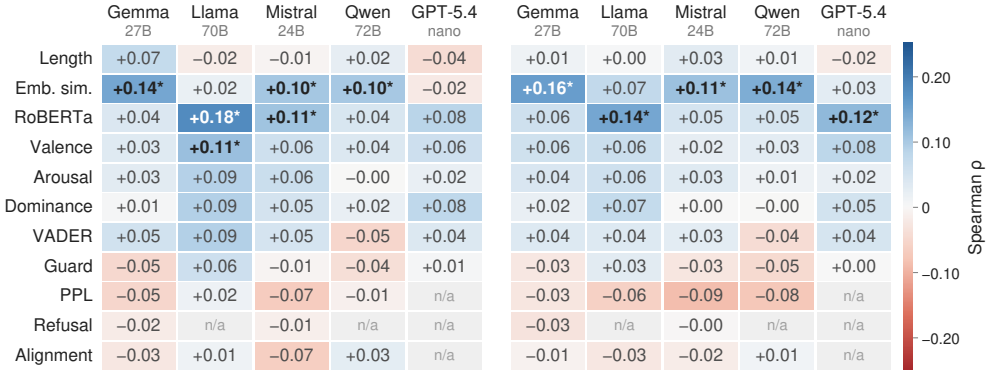
(f) Minimum, $PC_{sij}$

**Fig. D5**: *VP*: Spearman $\rho$ between $RS_{sij}$ / $PC_{sij}$ and the mean (a, b), maximum (c, d) and minimum (e, f) of each feature over the two instances (embedding similarity: computed on the pair). $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. 95% CIs are omitted for space.



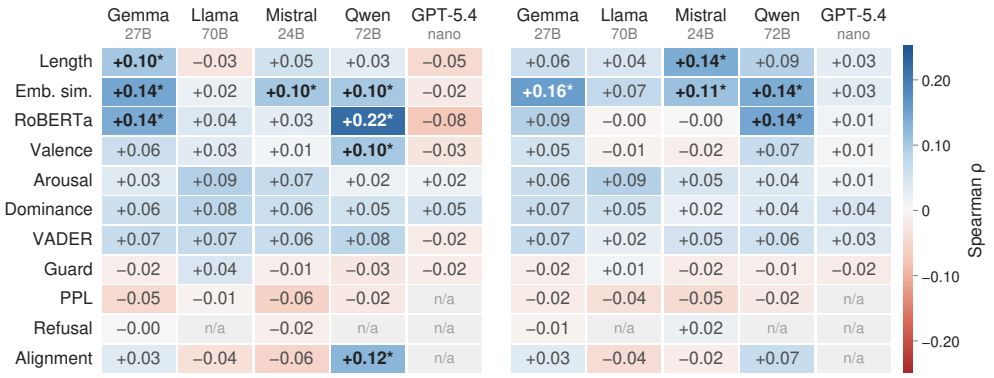

(a) Mean, $RS_{sij}$

(b) Mean, $PC_{sij}$




(c) Maximum, $RS_{sij}$

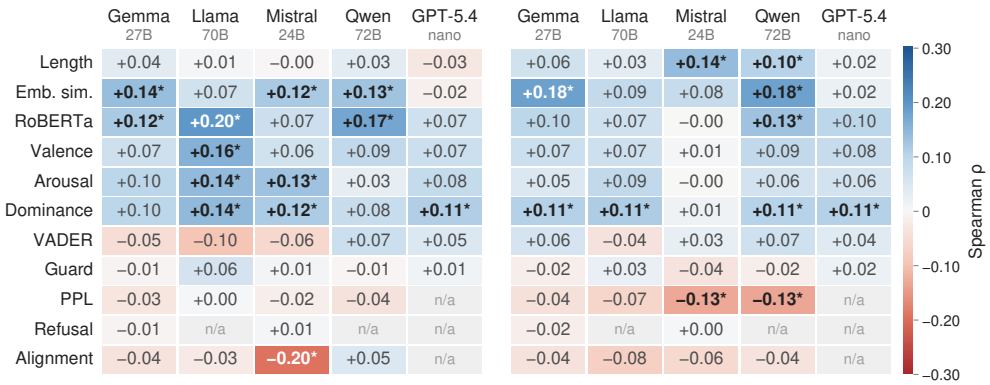
(d) Maximum, $PC_{sij}$




(e) Minimum, $RS_{sij}$

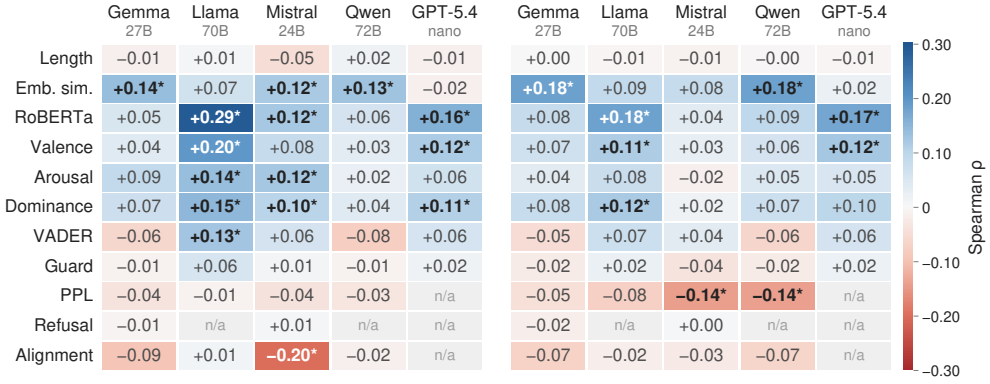
(f) Minimum, $PC_{sij}$

**Fig. D6**: *HB*: Spearman $\rho$ between $RS_{sij}$ / $PC_{sij}$ and the mean (a, b), maximum (c, d) and minimum (e, f) of each feature over the two instances (embedding similarity: computed on the pair). $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. 95% CIs are omitted for space.



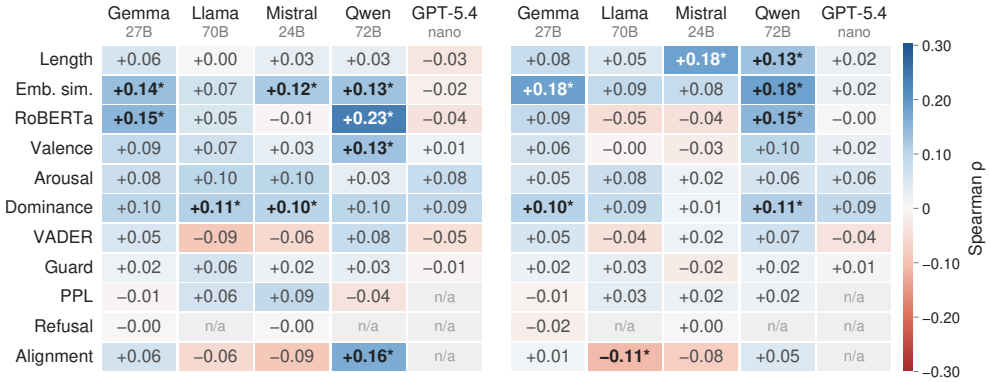

(a) Mean, $RS_{sij}$

(b) Mean, $PC_{sij}$




(c) Maximum, $RS_{sij}$

(d) Maximum, $PC_{sij}$



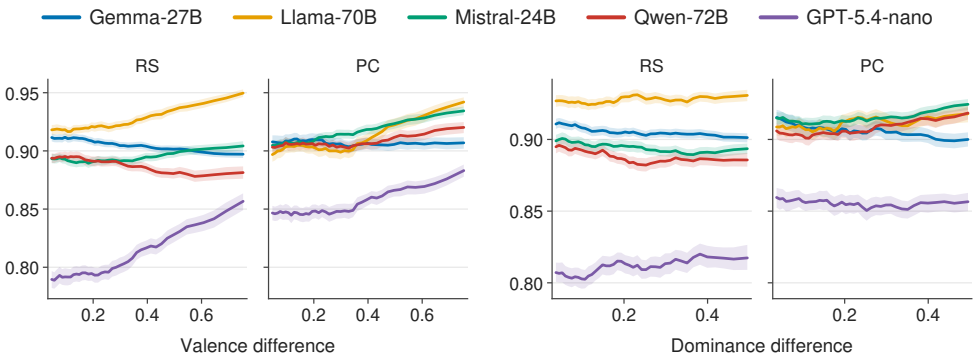

(e) Minimum, $RS_{sij}$

(f) Minimum, $PC_{sij}$

**Fig. D7**: *HBY*: Spearman $\rho$ between $RS_{sij}$ / $PC_{sij}$ and the mean (a, b), maximum (c, d) and minimum (e, f) of each feature over the two instances (embedding similarity: computed on the pair). $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. 95% CIs are omitted for space.

***Single-feature plots.***

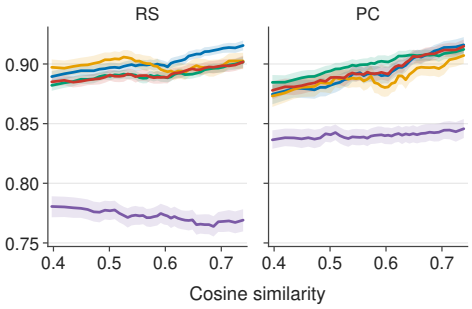
Figures D8 and D9 show $RS_{sij}$ and $PC_{sij}$ against the remaining features with at least one significant correlation in Figures 4, 5 and 6.



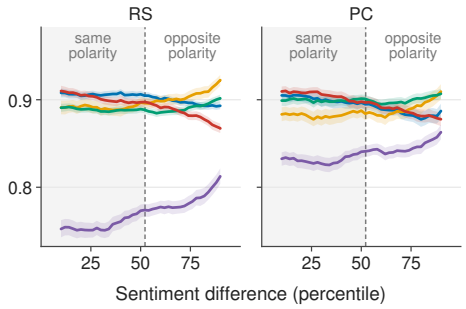

(a) VP: Difference in VAD-Valence.




(b) VP: Difference in VAD-Dominance.

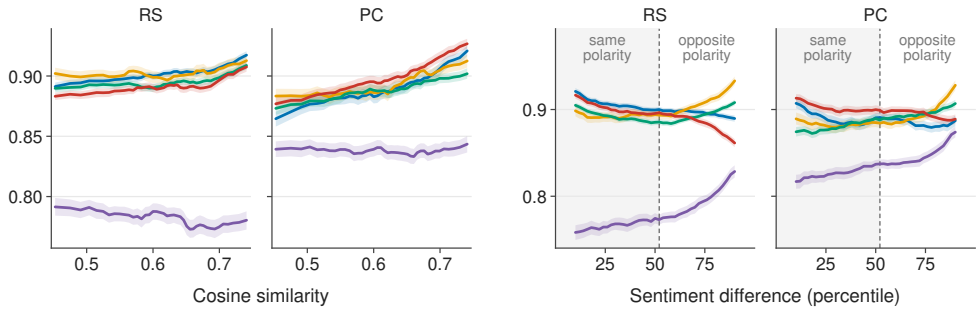



(c) HB: Cosine similarity between the texts.




(d) HB: Difference in Roberta-sentiment score.

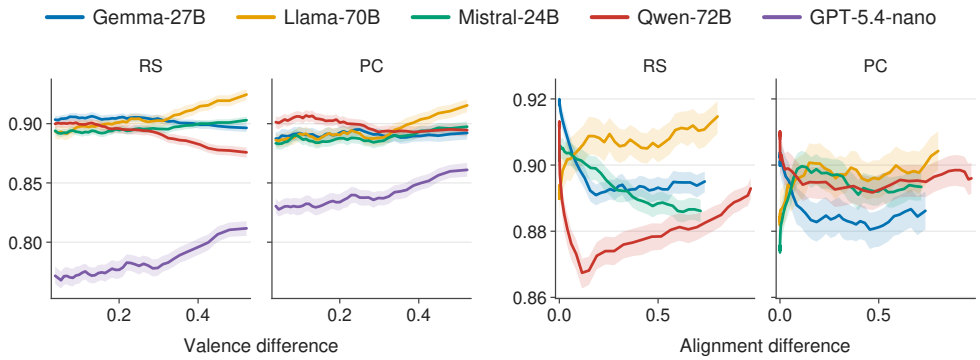
**Fig. D8**: $RS_{sij}$ and $PC_{sij}$ against the difference in feature between the two instances, for the features with at least one significant correlation not shown in Figure 7. Pairs are sorted by the feature and grouped into overlapping sliding windows of 20% of the pairs, shifted by 2% at each step. Each point is the average over a window, plotted at the window's mean feature value. Shaded areas are 95% bands from 1,000 bootstrap samples of situations. For Sentiment-Roberta, whose values are concentrated near 0 and 1, the x-axis shows the percentile of the feature, and the shaded region marks pairs whose opinions have the same polarity. (1/2)




(a) HBY: Cosine similarity between the texts.




(b) HBY: Difference in Roberta-sentiment score.




(c) HBY: Difference in VAD-Valence.




(d) HBY: Difference in alignment.

**Fig. D9**: $RS_{sij}$ and $PC_{sij}$ against the difference in feature between the two instances, for the features with at least one significant correlation not shown in Figure 7. Pairs are sorted by the feature and grouped into overlapping sliding windows of 20% of the pairs, shifted by 2% at each step. Each point is the average over a window, plotted at the window's mean feature value. Shaded areas are 95% bands from 1,000 bootstrap samples of situations. (2/2)

## D.3 Positional Bias Analyses

***Significance tables.***

Tables D4, D5 and D6 report the correlations of Figure 8 with their adjusted p-values.

| Feature | Judge | Pivot | $\rho$ [95% CI] | $q$ |
|---|---|---|---|---|
| Text length | Gemma-27B | Longer text | $-0.18^*$ [−0.21, −0.16] | 0.002 |
| | Llama-70B | Longer text | $-0.15^*$ [−0.17, −0.12] | 0.002 |
| | Mistral-24B | Longer text | $-0.22^*$ [−0.24, −0.20] | 0.002 |
| | Qwen-72B | Longer text | $-0.24^*$ [−0.27, −0.22] | 0.002 |
| | GPT-5.4-nano | Longer text | $-0.16^*$ [−0.19, −0.13] | 0.002 |
| Sentiment-Roberta | Gemma-27B | More positive | +0.01 [−0.01, +0.04] | 0.314 |
| | Llama-70B | More positive | +0.03 [+0.01, +0.05] | 0.015 |
| | Mistral-24B | More positive | +0.05 [+0.02, +0.08] | 0.003 |
| | Qwen-72B | More positive | −0.09 [−0.11, −0.06] | 0.003 |
| | GPT-5.4-nano | More positive | $+0.12^*$ [+0.09, +0.14] | 0.003 |
| VAD-Valence | Gemma-27B | Higher valence | +0.03 [+0.00, +0.06] | 0.041 |
| | Llama-70B | Higher valence | +0.01 [−0.02, +0.04] | 0.344 |
| | Mistral-24B | Higher valence | +0.02 [−0.01, +0.04] | 0.212 |
| | Qwen-72B | Higher valence | −0.07 [−0.09, −0.04] | 0.002 |
| | GPT-5.4-nano | Higher valence | $+0.12^*$ [+0.09, +0.14] | 0.002 |
| VAD-Arousal | Gemma-27B | Higher arousal | +0.08 [+0.06, +0.11] | 0.002 |
| | Llama-70B | Higher arousal | +0.05 [+0.03, +0.08] | 0.002 |
| | Mistral-24B | Higher arousal | +0.07 [+0.04, +0.10] | 0.002 |
| | Qwen-72B | Higher arousal | +0.09 [+0.07, +0.12] | 0.002 |
| | GPT-5.4-nano | Higher arousal | +0.08 [+0.05, +0.11] | 0.002 |
| VAD-Dominance | Gemma-27B | Higher dominance | $+0.10^*$ [+0.08, +0.13] | 0.002 |
| | Llama-70B | Higher dominance | +0.05 [+0.03, +0.08] | 0.002 |
| | Mistral-24B | Higher dominance | +0.08 [+0.06, +0.11] | 0.002 |
| | Qwen-72B | Higher dominance | +0.04 [+0.02, +0.07] | 0.002 |
| | GPT-5.4-nano | Higher dominance | $+0.15^*$ [+0.13, +0.18] | 0.002 |
| Sentiment-VADER | Gemma-27B | Higher compound | +0.02 [−0.01, +0.05] | 0.297 |
| | Llama-70B | Higher compound | +0.01 [−0.02, +0.04] | 0.394 |
| | Mistral-24B | Higher compound | +0.02 [−0.01, +0.04] | 0.297 |
| | Qwen-72B | Higher compound | −0.02 [−0.05, +0.01] | 0.297 |
| | GPT-5.4-nano | Higher compound | +0.04 [+0.01, +0.07] | 0.070 |
| Guard-Safety | Gemma-27B | Safer | +0.03 [−0.00, +0.06] | 0.137 |
| | Llama-70B | Safer | +0.04 [+0.01, +0.07] | 0.020 |
| | Mistral-24B | Safer | +0.00 [−0.02, +0.03] | 0.899 |
| | Qwen-72B | Safer | +0.08 [+0.05, +0.11] | 0.010 |
| | GPT-5.4-nano | Safer | −0.02 [−0.06, +0.01] | 0.142 |
| Perplexity | Gemma-27B | Lower PPL | −0.05 [−0.08, −0.03] | 0.003 |
| | Llama-70B | Lower PPL | −0.03 [−0.06, −0.01] | 0.020 |
| | Mistral-24B | Lower PPL | −0.07 [−0.10, −0.05] | 0.003 |
| | Qwen-72B | Lower PPL | −0.09 [−0.12, −0.06] | 0.003 |
| Refusal (SIT) | Gemma-27B | Less-refused stance | +0.09 [−0.10, +0.27] | 0.771 |
| | Llama-70B | Less-refused stance | −0.04 [−0.28, +0.23] | 0.969 |
| | Mistral-24B | Less-refused stance | +0.00 [−0.17, +0.19] | 0.969 |
| | Qwen-72B | Less-refused stance | −0.10 [−0.23, +0.04] | 0.771 |
| Alignment (SIT) | Gemma-27B | Majority stance | +0.02 [−0.03, +0.07] | 0.354 |
| | Llama-70B | Majority stance | −0.02 [−0.06, +0.02] | 0.354 |
| | Mistral-24B | Majority stance | −0.03 [−0.07, +0.01] | 0.308 |
| | Qwen-72B | Majority stance | $-0.11^*$ [−0.17, −0.06] | 0.008 |

**Table D4**: *VP*: Spearman correlation $\rho$ between the gap in each feature between the two instances and $\Delta_{sij}$ (score with the pivot first minus score with the pivot second), i.e. the values of Figure 8. 95% CI from situation-level bootstrap resamples; $q$: Benjamini–Hochberg within each feature (all its variants and judges); $^*$: $q < 0.05$ and $|\rho| \geq 0.1$.

| Feature | Judge | Pivot | $\rho$ [95% CI] | $q$ |
|---|---|---|---|---|
| Text length | Gemma-27B | Longer text | $-0.10^{*}$ [−0.13, −0.07] | 0.002 |
| | Llama-70B | Longer text | −0.08 [−0.11, −0.05] | 0.002 |
| | Mistral-24B | Longer text | $-0.13^{*}$ [−0.16, −0.10] | 0.002 |
| | Qwen-72B | Longer text | $-0.16^{*}$ [−0.19, −0.13] | 0.002 |
| | GPT-5.4-nano | Longer text | $-0.11^{*}$ [−0.13, −0.08] | 0.002 |
| Sentiment-Roberta | Gemma-27B | More positive | −0.08 [−0.11, −0.05] | 0.002 |
| | Llama-70B | More positive | −0.07 [−0.10, −0.05] | 0.002 |
| | Mistral-24B | More positive | −0.04 [−0.07, −0.01] | 0.022 |
| | Qwen-72B | More positive | $-0.21^{*}$ [−0.24, −0.18] | 0.002 |
| | GPT-5.4-nano | More positive | −0.08 [−0.11, −0.05] | 0.002 |
| VAD-Valence | Gemma-27B | Higher valence | −0.06 [−0.08, −0.03] | 0.010 |
| | Llama-70B | Higher valence | −0.07 [−0.10, −0.04] | 0.010 |
| | Mistral-24B | Higher valence | −0.01 [−0.04, +0.02] | 0.653 |
| | Qwen-72B | Higher valence | $-0.11^{*}$ [−0.14, −0.08] | 0.010 |
| | GPT-5.4-nano | Higher valence | −0.05 [−0.08, −0.02] | 0.015 |
| VAD-Arousal | Gemma-27B | Higher arousal | +0.03 [+0.00, +0.06] | 0.069 |
| | Llama-70B | Higher arousal | +0.01 [−0.02, +0.03] | 0.612 |
| | Mistral-24B | Higher arousal | +0.02 [−0.01, +0.05] | 0.332 |
| | Qwen-72B | Higher arousal | +0.03 [+0.00, +0.06] | 0.069 |
| | GPT-5.4-nano | Higher arousal | +0.02 [−0.01, +0.05] | 0.251 |
| VAD-Dominance | Gemma-27B | Higher dominance | −0.02 [−0.05, +0.02] | 0.332 |
| | Llama-70B | Higher dominance | −0.03 [−0.05, +0.00] | 0.120 |
| | Mistral-24B | Higher dominance | −0.02 [−0.05, +0.01] | 0.220 |
| | Qwen-72B | Higher dominance | −0.03 [−0.07, −0.00] | 0.069 |
| | GPT-5.4-nano | Higher dominance | −0.02 [−0.05, +0.01] | 0.228 |
| Sentiment-VADER | Gemma-27B | Higher compound | −0.03 [−0.06, +0.00] | 0.075 |
| | Llama-70B | Higher compound | −0.04 [−0.07, −0.01] | 0.035 |
| | Mistral-24B | Higher compound | −0.01 [−0.04, +0.02] | 0.619 |
| | Qwen-72B | Higher compound | −0.08 [−0.11, −0.05] | 0.010 |
| | GPT-5.4-nano | Higher compound | −0.03 [−0.06, −0.00] | 0.063 |
| Guard-Safety | Gemma-27B | Safer | −0.02 [−0.05, +0.02] | 0.352 |
| | Llama-70B | Safer | −0.03 [−0.06, −0.00] | 0.050 |
| | Mistral-24B | Safer | −0.05 [−0.08, −0.02] | 0.010 |
| | Qwen-72B | Safer | −0.03 [−0.07, +0.00] | 0.092 |
| | GPT-5.4-nano | Safer | −0.05 [−0.08, −0.02] | 0.015 |
| Perplexity | Gemma-27B | Lower PPL | −0.08 [−0.10, −0.05] | 0.002 |
| | Llama-70B | Lower PPL | −0.09 [−0.12, −0.07] | 0.002 |
| | Mistral-24B | Lower PPL | $-0.11^{*}$ [−0.13, −0.08] | 0.002 |
| | Qwen-72B | Lower PPL | $-0.18^{*}$ [−0.21, −0.15] | 0.002 |
| Alignment (SIT) | Gemma-27B | Majority stance | +0.01 [−0.03, +0.04] | 0.858 |
| | Llama-70B | Majority stance | −0.00 [−0.04, +0.03] | 0.945 |
| | Mistral-24B | Majority stance | −0.01 [−0.05, +0.03] | 0.858 |
| | Qwen-72B | Majority stance | +0.06 [+0.02, +0.11] | 0.008 |

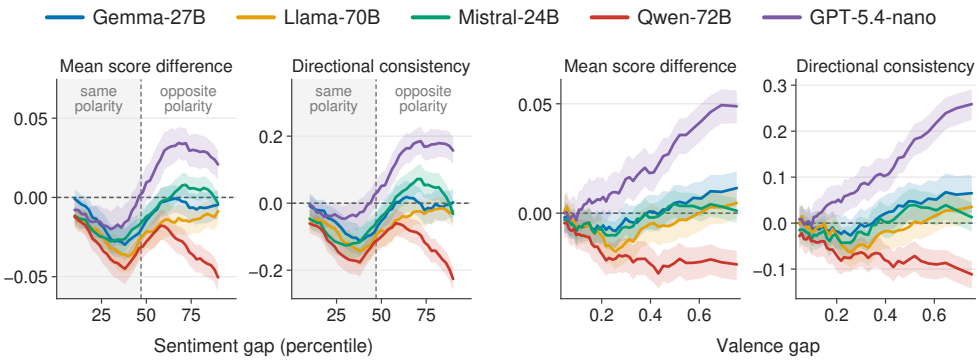
**Table D5**: *HB*: Spearman correlation $\rho$ between the gap in each feature between the two instances and $\Delta_{sij}$ (score with the pivot first minus score with the pivot second), i.e. the values of Figure 8. 95% CI from situation-level bootstrap resamples; $q$: Benjamini–Hochberg within each feature (all its variants and judges); $^{*}$: $q < 0.05$ and $|\rho| \geq 0.1$.

| Feature | Judge | Pivot | $\rho$ [95% CI] | $q$ |
|---|---|---|---|---|
| Text length | `Gemma-27B` | Longer text | $-0.10^{*}$ [−0.12, −0.08] | 0.002 |
| | `Llama-70B` | Longer text | −0.09 [−0.11, −0.07] | 0.002 |
| | `Mistral-24B` | Longer text | $-0.18^{*}$ [−0.20, −0.16] | 0.002 |
| | `Qwen-72B` | Longer text | $-0.18^{*}$ [−0.20, −0.15] | 0.002 |
| | `GPT-5.4-nano` | Longer text | $-0.13^{*}$ [−0.15, −0.11] | 0.002 |
| Sentiment-Roberta | `Gemma-27B` | More positive | +0.01 [−0.01, +0.03] | 0.492 |
| | `Llama-70B` | More positive | −0.01 [−0.03, +0.01] | 0.240 |
| | `Mistral-24B` | More positive | +0.08 [+0.06, +0.10] | 0.005 |
| | `Qwen-72B` | More positive | $-0.13^{*}$ [−0.16, −0.10] | 0.005 |
| | `GPT-5.4-nano` | More positive | +0.03 [+0.00, +0.05] | 0.070 |
| VAD-Valence | `Gemma-27B` | Higher valence | −0.03 [−0.05, −0.00] | 0.095 |
| | `Llama-70B` | Higher valence | −0.05 [−0.08, −0.03] | 0.006 |
| | `Mistral-24B` | Higher valence | +0.01 [−0.01, +0.03] | 0.548 |
| | `Qwen-72B` | Higher valence | −0.09 [−0.11, −0.06] | 0.006 |
| | `GPT-5.4-nano` | Higher valence | −0.00 [−0.03, +0.02] | 0.729 |
| VAD-Arousal | `Gemma-27B` | Higher arousal | −0.01 [−0.03, +0.02] | 0.666 |
| | `Llama-70B` | Higher arousal | −0.01 [−0.03, +0.01] | 0.548 |
| | `Mistral-24B` | Higher arousal | −0.05 [−0.07, −0.03] | 0.006 |
| | `Qwen-72B` | Higher arousal | −0.02 [−0.04, +0.01] | 0.338 |
| | `GPT-5.4-nano` | Higher arousal | −0.03 [−0.06, −0.01] | 0.006 |
| VAD-Dominance | `Gemma-27B` | Higher dominance | +0.01 [−0.01, +0.04] | 0.472 |
| | `Llama-70B` | Higher dominance | +0.01 [−0.02, +0.03] | 0.666 |
| | `Mistral-24B` | Higher dominance | +0.06 [+0.04, +0.09] | 0.006 |
| | `Qwen-72B` | Higher dominance | +0.01 [−0.02, +0.03] | 0.666 |
| | `GPT-5.4-nano` | Higher dominance | +0.01 [−0.01, +0.04] | 0.513 |
| Sentiment-VADER | `Gemma-27B` | Higher compound | −0.05 [−0.08, −0.03] | 0.002 |
| | `Llama-70B` | Higher compound | −0.04 [−0.06, −0.02] | 0.002 |
| | `Mistral-24B` | Higher compound | −0.05 [−0.08, −0.03] | 0.002 |
| | `Qwen-72B` | Higher compound | $-0.11^{*}$ [−0.14, −0.09] | 0.002 |
| | `GPT-5.4-nano` | Higher compound | −0.04 [−0.07, −0.02] | 0.002 |
| Guard-Safety | `Gemma-27B` | Safer | −0.06 [−0.08, −0.03] | 0.002 |
| | `Llama-70B` | Safer | −0.05 [−0.07, −0.03] | 0.002 |
| | `Mistral-24B` | Safer | $-0.12^{*}$ [−0.14, −0.09] | 0.002 |
| | `Qwen-72B` | Safer | −0.08 [−0.10, −0.05] | 0.002 |
| | `GPT-5.4-nano` | Safer | $-0.10^{*}$ [−0.12, −0.08] | 0.002 |
| Perplexity | `Gemma-27B` | Lower PPL | −0.07 [−0.09, −0.05] | 0.002 |
| | `Llama-70B` | Lower PPL | −0.09 [−0.12, −0.07] | 0.002 |
| | `Mistral-24B` | Lower PPL | $-0.12^{*}$ [−0.13, −0.10] | 0.002 |
| | `Qwen-72B` | Lower PPL | $-0.20^{*}$ [−0.23, −0.18] | 0.002 |
| Alignment (SIT) | `Gemma-27B` | Majority stance | +0.00 [−0.03, +0.03] | 0.849 |
| | `Llama-70B` | Majority stance | −0.01 [−0.04, +0.02] | 0.578 |
| | `Mistral-24B` | Majority stance | −0.10 [−0.12, −0.07] | 0.004 |
| | `Qwen-72B` | Majority stance | −0.07 [−0.10, −0.05] | 0.004 |

**Table D6**: $HBY$: Spearman correlation $\rho$ between the gap in each feature between the two instances and $\Delta_{sij}$ (score with the pivot first minus score with the pivot second), i.e. the values of Figure 8. 95% CI from situation-level bootstrap resamples; $q$: Benjamini–Hochberg within each feature (all its variants and judges); $^{*}$: $q < 0.05$ and $|\rho| \geq 0.1$.

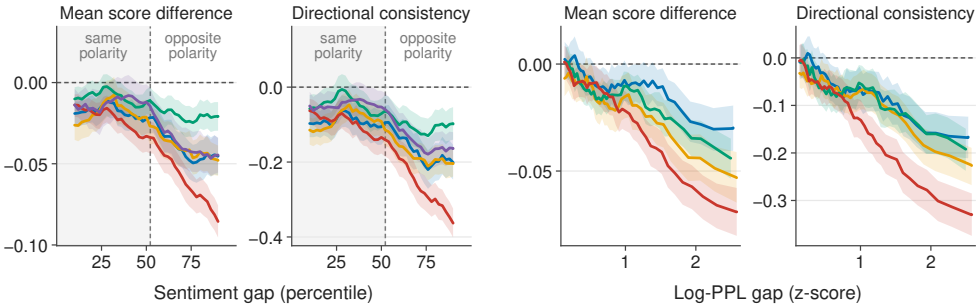
***Single-feature plots.***

Figures D10 and D11 show the mean oriented score difference and the directional consistency against the remaining features with at least one significant correlation.




(a) VP: Gap in Roberta-sentiment score (pivot: more positive opinion).

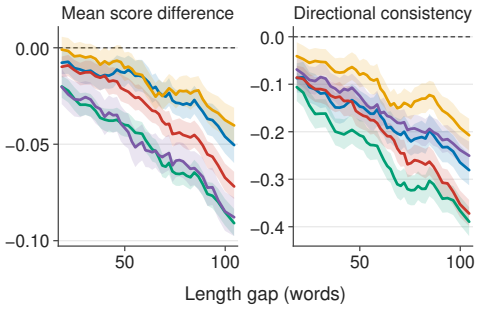
(b) VP: Gap in VAD-Valence (pivot: more positive opinion).



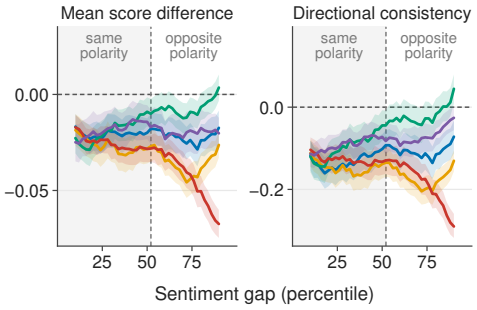

(c) HB: Gap in Roberta-sentiment score (pivot: more positive opinion).

(d) HB: Gap in perplexity (pivot: lower-perplexity opinion).

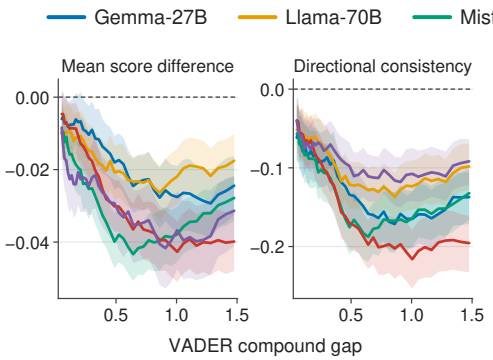
**Fig. D10**: Mean oriented score difference $\overline{\Delta}$ (left panel of each plot) and directional consistency $\overline{\text{sign}(\Delta)}$ (right panel) against the gap in a feature between the two instances, for the features with at least one significant correlation not shown in Figure 9. $\Delta_{sij}$ is oriented on the pivot, so negative values mean a lower score when the pivot is presented first than when it is presented second; the dashed line marks the absence of directional bias. Pairs are sorted by the feature and grouped into overlapping sliding windows of 20% of the pairs, shifted by 2% at each step. Each point is the average over a window, plotted at the window's mean feature value. Shaded areas are 95% bands from 1,000 bootstrap samples of situations. For Sentiment-Roberta, whose values are concentrated near 0 and 1, the x-axis shows the percentile of the feature, and the shaded region marks pairs whose opinions have the same polarity. (1/2)



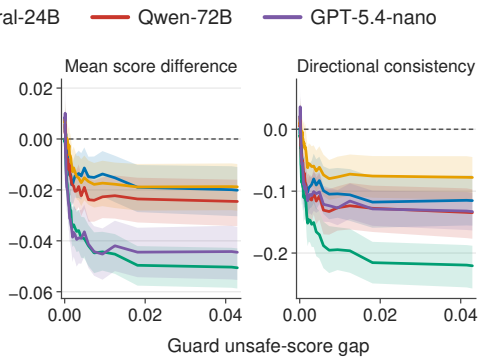

(a) HBY: Gap in text length (pivot: longer opinion).




(b) HBY: Gap in Roberta-sentiment score (pivot: more positive opinion).




(c) HBY: Gap in VADER compound score (pivot: more positive opinion).




(d) HBY: Gap in Guard unsafe score (pivot: safer opinion).

**Fig. D11**: Mean oriented score difference $\overline{\Delta}$ (left panel of each plot) and directional consistency $\overline{\text{sign}(\Delta)}$ (right panel) against the gap in a feature between the two instances, for the features with at least one significant correlation not shown in Figure 9. $\Delta_{sij}$ is oriented on the pivot, so negative values mean a lower score when the pivot is presented first than when it is presented second; the dashed line marks the absence of directional bias. Pairs are sorted by the feature and grouped into overlapping sliding windows of 20% of the pairs, shifted by 2% at each step. Each point is the average over a window, plotted at the window's mean feature value. Shaded areas are 95% bands from 1,000 bootstrap samples of situations. (2/2)

## D.4 Joint Analysis of Correlated Features

The analyses in the main text test one feature at a time, but several features are correlated. The sentiment and affect scores measure related properties of the text: the pairwise Spearman correlations between the gaps in Sentiment-RoBERTa, VAD-Valence, VAD-Dominance and Sentiment-VADER range from 0.25 to 0.66. Longer opinions also have lower perplexity: depending on the judge, the correlation between the gaps in length and perplexity is 0.41–0.81 on Habermas, 0.38–0.75 on Habermas-Hybrid and 0.26–0.41 on ValuePrism. A feature could therefore appear significant only because it is correlated with another one. To test this, we model all features jointly for each dataset and judge.

***Method.***

For stability, we compute the partial Spearman correlation between each feature (the gap between the two opinions, or embedding similarity) and $RS_{sij}$ or $PC_{sij}$, controlling for all other features. For directional bias, we regress $\Delta_{sij}$ on the gaps of all features at once. Each gap is oriented towards its pivot and divided by its root mean square, so that its coefficient is the change in $\Delta_{sij}$ for a typical gap. The model has no intercept, so it is unchanged when the two opinions are swapped. In both cases, 95% CIs and p-values come from 1,000 situation-level bootstrap resamples. P-values are corrected with Benjamini–Hochberg within each dataset, separately for stability (jointly over all features, judges and both metrics) and for directional bias (jointly over all features and judges). Unlike in the main analysis, where the correction is applied within each feature, here the family includes all features, which makes the joint test more conservative.

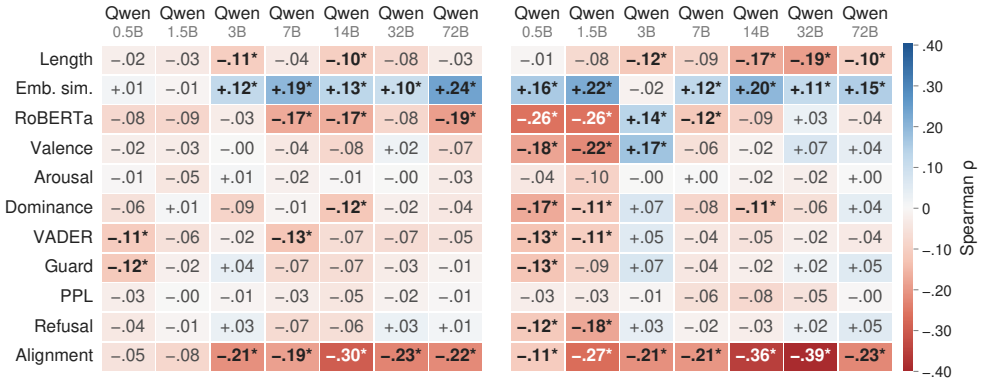
***Results.***

Table D7 reports, for each association marked as significant in the main text, whether it remains significant with the same sign in the joint model. For stability, 63 of 67 do, and 49 of these also keep $|\rho| \geq 0.1$. For directional bias, 23 of 28 do. The nine associations that do not remain are: (i) Sentiment-RoBERTa, VAD-Valence and VAD-Dominance for `Gemma-27B` (stability, ValuePrism), whose effect is shared with the other affect features and with embedding similarity, which correlates with the affect gaps on ValuePrism ($\rho$ between $-0.3$ and $-0.4$); (ii) Sentiment-RoBERTa and VAD-Valence for `GPT-5.4-nano` (directional bias, ValuePrism), where VAD-Dominance remains significant; (iii) text length for `Qwen-72B` ($PC_{sij}$, Habermas-Hybrid), where perplexity remains significant; and (iv) Guard-Safety (`Mistral-24B`, `GPT-5.4-nano`) and Sentiment-VADER (`Qwen-72B`) in directional bias on Habermas-Hybrid. A few affect coefficients that are not significant in the main text become significant with the opposite sign in the joint models (e.g. VAD-Arousal in directional bias on Habermas-Hybrid). We attribute these to the correlation among the affect features and do not interpret them.

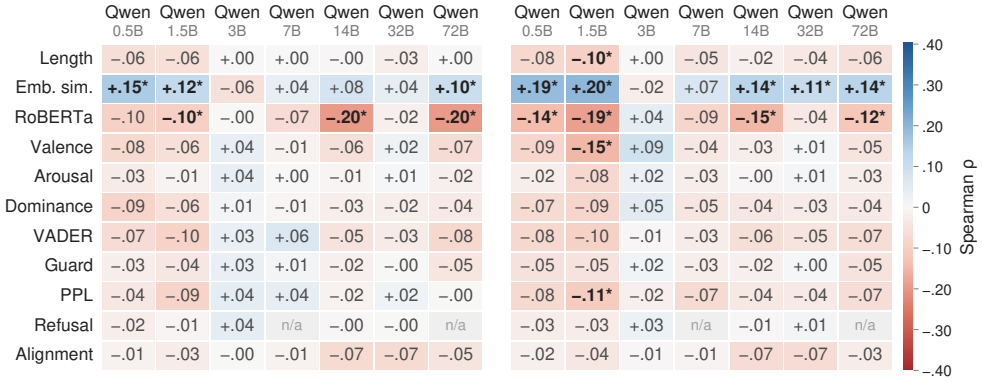
| | Stability | | | Positional bias | | |
|---|---|---|---|---|---|---|
| Feature | VP | HB | HBY | VP | HB | HBY |
| Length | 8/8 | – | 1/2 | 5/5 | 4/4 | 4/4 |
| Emb. sim. | 5/5 | 6/6 | 5/5 | – | – | – |
| RoBERTa | 6/7 | 5/5 | 7/7 | 0/1 | 1/1 | 1/1 |
| Valence | 5/6 | – | 2/2 | 0/1 | 1/1 | – |
| Dominance | 0/1 | – | – | 2/2 | – | – |
| VADER | – | – | – | – | – | 0/1 |
| Guard | – | – | – | – | – | 0/2 |
| PPL | – | – | 2/2 | – | 2/2 | 2/2 |
| Alignment | 7/7 | – | 4/4 | 1/1 | – | – |
| Total | 31/34 | 11/11 | 21/22 | 8/10 | 8/8 | 7/10 |

**Table D7**: Associations marked as significant in the main text (over judges and, for stability, both $RS_{sij}$ and $PC_{sij}$) that remain significant with the same sign when all features are modelled jointly, as kept/marked. –: no association marked in the main text.

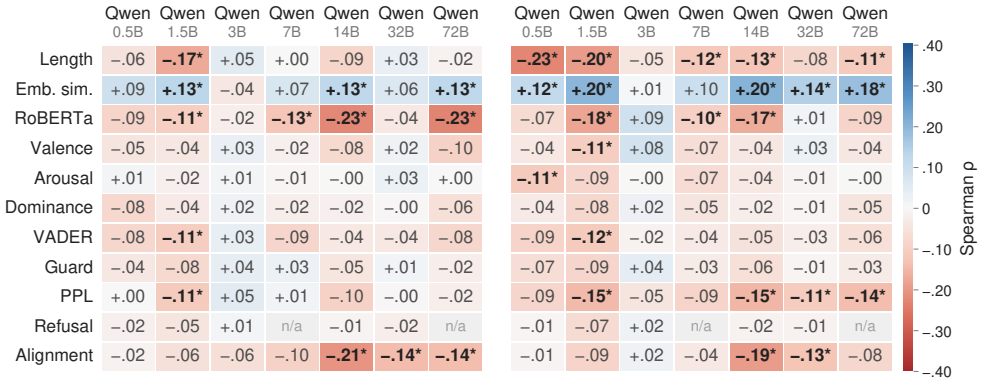
## D.5 Model Scale Analyses




(a) VP, $RS_{sij}$ (b) VP, $PC_{sij}$

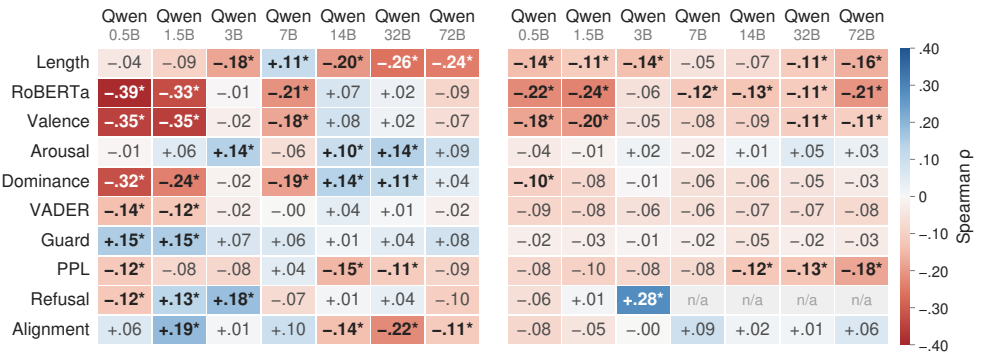



(c) HB, $RS_{sij}$ (d) HB, $PC_{sij}$




(e) HBY, $RS_{sij}$ (f) HBY, $PC_{sij}$

**Fig. D12**: Model scale: Spearman $\rho$ between $RS_{sij}$ / $PC_{sij}$ and the difference in each feature between the two instances, for the Qwen2.5 family on *VP* (a, b), *HB* (c, d) and *HBY* (e, f). $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. 95% CIs are omitted for space.




(a) *VP* (b) *HB*

| | Qwen 0.5B | Qwen 1.5B | Qwen 3B | Qwen 7B | Qwen 14B | Qwen 32B | Qwen 72B |
|---|---|---|---|---|---|---|---|
| Length | **−.28*** | **−.14*** | **−.16*** | −.00 | **−.14*** | **−.14*** | **−.18*** |
| RoBERTa | **−.16*** | **−.24*** | +.01 | **−.11*** | **−.10*** | +.00 | **−.13*** |
| Valence | **−.14*** | **−.21*** | −.04 | **−.13*** | −.06 | −.02 | −.09 |
| Arousal | **−.22*** | **−.17*** | +.00 | **−.10*** | −.03 | +.04 | −.02 |
| Dominance | −.06 | **−.12*** | −.01 | **−.13*** | −.03 | +.03 | +.01 |
| VADER | **−.15*** | **−.16*** | −.06 | −.04 | −.10 | −.07 | **−.11*** |
| Guard | −.10 | −.07 | −.07 | +.03 | −.08 | −.08 | −.08 |
| PPL | **−.12*** | **−.11*** | **−.16*** | +.02 | **−.22*** | **−.16*** | **−.20*** |
| Refusal | −.07 | −.01 | +.12 | n/a | n/a | n/a | n/a |
| Alignment | −.09 | −.08 | +.01 | **+.14*** | −.03 | **−.14*** | −.07 |

(c) *HBY*

**Fig. D13**: Model scale: Spearman $\rho$ between $\Delta_{sij}$ and the gap in each feature between the two instances, for the Qwen2.5 family. $*$: adjusted p-value $< 0.05$ (correction within feature) and $|\rho| \geq 0.1$. 95% CIs are omitted for space.

# References


[1] Sai, A.B., Mohankumar, A.K., Khapra, M.M.: A survey of evaluation metrics used for nlg systems. ACM Computing Surveys (CSUR) **55**, 1–39 (2020)

[2] Sai, A.B., Dixit, T., Sheth, D.Y., Mohan, S., Khapra, M.M.: Perturbation CheckLists for evaluating NLG evaluation metrics. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7219–7234. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.575 . https://aclanthology.org/2021.emnlp-main.575/

[3] Schmidtova, P., Mahamood, S., Balloccu, S., Dusek, O., Gatt, A., Gkatzia, D.,

Howcroft, D.M., Platek, O., Sivaprasad, A.: Automatic metrics in natural language generation: A survey of current evaluation practices. In: Mahamood, S., Minh, N.L., Ippolito, D. (eds.) Proceedings of the 17th International Natural Language Generation Conference, pp. 557–583. Association for Computational Linguistics, Tokyo, Japan (2024). https://doi.org/10.18653/v1/2024.inlg-main.44 . https://aclanthology.org/2024.inlg-main.44/

[4] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23. Curran Associates Inc., Red Hook, NY, USA (2023)

[5] Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., Zhou, J.: Is ChatGPT a good NLG evaluator? a preliminary study. In: Dong, Y., Xiao, W., Wang, L., Liu, F., Carenini, G. (eds.) Proceedings of the 4th New Frontiers in Summarization Workshop, pp. 1–11. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.newsum-1.1 . https://aclanthology.org/2023.newsum-1.1/

[6] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: NLG evaluation using gpt-4 with better human alignment. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 2511–2522. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.153 . https://aclanthology.org/2023.emnlp-main.153/

[7] Li, Z., Wang, C., Ma, P., Wu, D., Wang, S., Gao, C., Liu, Y.: Split and merge: Aligning position biases in LLM-based evaluators. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 11084–11108. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.621 . https://aclanthology.org/2024.emnlp-main.621/

[8] Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., Chawla, N.V., Zhang, X.: Justice or prejudice? quantifying biases in LLM-as-a-judge. In: The Thirteenth International Conference on Learning Representations (2025). https://openreview.net/forum?id=3GTtZFiajM

[9] Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., Prakash, S.: Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In: International Conference on Machine Learning (2023). https://api.semanticscholar.org/CorpusID:261493811

[10] Li, D., Yang, S., Tan, Z., Baik, J.Y., Yun, S.-Y., Lee, J., Chacko, A., Hou, B., Duong-Tran, D.A., Ding, Y., Liu, H., Shen, L., Chen, T.: Dalk: Dynamic

co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing **2024**, 2187–2205 (2024)

[11] Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., Tu, Z.: Encouraging divergent thinking in large language models through multi-agent debate. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 17889–17904. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.992 . https://aclanthology.org/2024.emnlp-main.992/

[12] Wu, S., Choi, E., Khatua, A., Wang, Z., He-Yueya, J., Weerasooriya, T.C., Wei, W., Yang, D., Leskovec, J., Zou, J.: HumanLM: Simulating users with state alignment beats response imitation. In: Forty-third International Conference on Machine Learning (2026). https://openreview.net/forum?id=xyFVjBi9wC

[13] Zhou, X., Sun, W., Du, W., Liu, J., Sun, H., Ma, Q., Wu, T., Yang, Y., Sap, M.: OdysSim: Building Foundation Models for Human Behavior Simulation. arXiv. arXiv:2606.14199 [cs.CL] (2026). https://doi.org/10.48550/arXiv.2606.14199 . http://arxiv.org/abs/2606.14199 Accessed 2026-06-17

[14] Wang, X., Wang, H., Zhang, Y., Yuan, X., Xu, R., Huang, J.-t., Yuan, S., Guo, H., Chen, J., Zhou, S., Wang, W., Xiao, Y.: Coser: coordinating llm-based persona simulation of established roles. In: Proceedings of the 42nd International Conference on Machine Learning. ICML'25, pp. 1–37. JMLR.org, Vancouver, Canada (2025)

[15] Dou, Y., Galley, M., Peng, B., Kedzie, C., Cai, W., Ritter, A., Quirk, C., Xu, W., Gao, J.: SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 35212–35290. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.18653/v1/2025.emnlp-main.1786 . https://aclanthology.org/2025.emnlp-main.1786/

[16] Hathidara, A., Yu, J., Senthil, V., Schreiber, S., Ankisettipalli, A.B.: Mirrorbench: A benchmark to evaluate conversational user-proxy agents for human-likeness. In: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. KDD '26, pp. 8978–8988. Association for Computing Machinery, New York, NY, USA (2026). https://doi.org/10.1145/3770855.3817496 . https://doi.org/10.1145/3770855.3817496

[17] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? In: International Conference on Machine Learning, pp. 29971–30004 (2023). PMLR

[18] Zhou, K., Constantinides, M., Quercia, D.: Should LLMs Be WEIRD? Exploring WEIRDness and Human Rights in Large Language Models. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society **8**(3), 2808–2820 (2025) https://doi.org/10.1609/aies.v8i3.36761 . Accessed 2026-09-04

[19] Mohammadi, H., Bagheri, R.: Exploring cultural variations in moral judgments with large language models. Computational Linguistics in the Netherlands Journal **15**, 59–77 (2026). Publisher Copyright: © 2026 Hadi Mohammadi & Robert A. Bagheri.; 35th Computational Linguistics in the Netherlands Conference, CLIN 2026 ; Conference date: 12-09-2025 Through 12-09-2025

[20] Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D.W., Pan, J., Tsvetkov, Y., Reinecke, K.: Biased LLMs can Influence Political Decision-Making. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6559–6607. Association for Computational Linguistics, Vienna, Austria (2025). https://aclanthology.org/2025.acl-long.328/ Accessed 2025-07-23

[21] Hackenburg, K., Tappin, B.M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D.G., Summerfield, C.: The levers of political persuasion with conversational artificial intelligence. Science **390**(6777), 3884 (2025) https://doi.org/10.1126/science.aea3884 https://www.science.org/doi/pdf/10.1126/science.aea3884

[22] Lin, H., Czarnek, G., Lewis, B., White, J.P., Berinsky, A.J., Costello, T., Pennycook, G., Rand, D.G.: Persuading voters using human–artificial intelligence dialogues. Nature (2025) https://doi.org/10.1038/s41586-025-09771-9

[23] Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., Vosoughi, S.: Judging the judges: A systematic study of position bias in LLM-as-a-judge. In: Inui, K., Sakti, S., Wang, H., Wong, D.F., Bhattacharyya, P., Banerjee, B., Ekbal, A., Chakraborty, T., Singh, D.P. (eds.) Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pp. 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai, India (2025). https://doi.org/10.18653/v1/2025.ijcnlp-long.18 . https://aclanthology.org/2025.ijcnlp-long.18/

[24] Brown, T., Mann, B., ...: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)

[25] Wei, J., *et al.*: Finetuned language models are zero-shot learners. In: International Conference on Learning Representations (2022)

[26] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al.: A survey on llm-as-a-judge. The Innovation **7**(6) (2026)

[27] Wataoka, K., Takahashi, T., Ri, R.: Self-preference bias in LLM-as-a-judge. In: Neurips Safe Generative AI Workshop 2024 (2024). https://openreview.net/forum?id=tLZZZIgPJX

[28] Panickssery, A., Bowman, S.R., Feng, S.: Llm evaluators recognize and favor their own generations. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. NIPS '24. Curran Associates Inc., Red Hook, NY, USA (2024)

[29] Li, B.-W., Wu, W., Tang, Z., Shi, L., Yang, J., Li, J.-Y., Yao, S., Qian, C., Hui, B., Zhang, Q., Yu, Z.-Y., Du, H., Yang, P., Lin, D.-H., Peng, C., Chen, K.: Devbench: A comprehensive benchmark for software development. ArXiv **abs/2403.08604** (2024)

[30] Wang, V., Zhang, M.J., Choi, E.: Improving llm-as-a-judge inference with the judgment distribution. arXiv preprint arXiv:2503.03064 (2025)

[31] Nussbaum, Z., Duderstadt, B.: Training Sparse Mixture Of Experts Text Embedding Models (2025). https://arxiv.org/abs/2502.07972

[32] Hartmann, J., Heitmann, M., Siebert, C., Schamp, C.: More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing **40**(1), 75–87 (2023) https://doi.org/10.1016/j.ijresmar.2022.05.005

[33] Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. (2015)

[34] Llama Team, A..M.: The Llama 3 Herd of Models (2024). https://arxiv.org/abs/2407.21783

[35] Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., Nanda, N.: Refusal in language models is mediated by a single direction. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems, vol. 37, pp. 136037–136083. Curran Associates, Inc., Red Hook, NY, USA (2024). https://doi.org/10.52202/079017-4322

[36] Scherrer, N., Shi, C., Feder, A., Blei, D.: Evaluating the moral beliefs encoded in llms. In: Thirty-seventh Conference on Neural Information Processing Systems (2023). https://openreview.net/forum?id=O06z2G18me

[37] Sorensen, T., Jiang, L., Hwang, J.D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., Choi, Y.: Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. Proceedings of the AAAI Conference on Artificial Intelligence **38**(18), 19937–19947 (2024) https://doi.org/10.1609/aaai.v38i18.29970 . Accessed 2025-09-03

[38] Tessler, M.H., Bakker, M.A., Jarrett, D., Sheahan, H., Chadwick, M.J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D.C., Botvinick, M., Summerfield, C.: AI can help humans find common ground in democratic deliberation. Science **386**(6719), 2852 (2024) https://doi.org/10.1126/science.adq2852 . Accessed 2025-09-03

[39] Kamath, G.T.A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-B., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X.-H., Tsitsulin, A., Busa-Fekete, R.I., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A.P., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., Gyorgy, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D.Z., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Pluci'nska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J.F., Lai, J., Orbay, J., Fernandez, J., Newlan, J., Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P.-m., Stańczyk, P., Tafti, P.D., Shivanna, R., Wu, R., Pan, R., Rokni, R.A., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V.N., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V.S., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Lepikhin, D., Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report. ArXiv **abs/2503.19786** (2025)

[40] Mistral AI: Mistral-Small-3.2-24B-Instruct-2506. https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506. Large language model (2025)

[41] Yang, Q.A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B.-W., Li, C.-Y., Liu, D., Huang, F., Dong, G., Wei, H.-R., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J.-X.,

Yang, J.-X., Zhou, J.-R., Lin, J.-Y., Dang, K., Lu, K.-M., Bao, K.-Q., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R.-J., Li, T., Xia, T., Ren, X.-Z., Ren, X.-C., Fan, Y., Su, Y., Zhang, Y.-C., Wan, Y., Liu, Y., Cui, Z., Zhang, Z.-R., Qiu, Z.-H., Quan, S., Wang, Z.: Qwen2.5 technical report. ArXiv **abs/2412.15115** (2024)

[42] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 289–300 (1995). Accessed 2026-09-28

[43] Lin, M., Lucas, H.C., Shmueli, G.: Research Commentary: Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research **24**(4), 906–917 (2013). Accessed 2026-06-22

[44] Chiang, C.-H., Lee, H.-y.: Can large language models be an alternative to human evaluations? In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15607–15631. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.870 . https://aclanthology.org/2023.acl-long.870/

[45] Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., Goldberg, Y.: Measuring and improving consistency in pretrained language models. Transactions of the Association for Computational Linguistics **9**, 1012–1031 (2021) https://doi.org/10.1162/tacl_a_00410

[46] Artstein, R., Poesio, M.: Survey article: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008) https://doi.org/10.1162/coli.07-034-R2

[47] Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., Demartini, G.: Let's agree to disagree: Fixing agreement measures for crowdsourcing. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 5, pp. 11–20 (2017)