Label and Explanation Variation in LLM-Based Annotation: a Case Study in Natural Language Inference

Artur Kulmizev[1] **Erika Lombart**[1] **Patrick Watrin**[1] **Marie-Catherine de Marneffe**[1,2]
[1]CENTAL, UCLouvain [2]FNRS
first.last@uclouvain.be

Abstract

Large language models (LLMs) have shown considerable promise for annotation purposes, yet questions remain about their ability to capture human label variation (HLV) — genuine disagreement between annotators often observed across NLP tasks. Here, we investigate how label and explanation variation manifests within *and* across LLMs with respect to the Natural Language Inference (NLI) task in English. Using zero-shot prompting with exact human annotation instructions, we treat individual model generations as *participants* and examine three response sampling strategies: varying generation parameters, leveraging within-family model size differences, and pooling responses from distinct LLMs. We show that, while model ensembles can generate label distributions similar to humans, they likewise exhibit distinct, idiosyncratic judgments and disagreement patterns. We further analyze explanation variation, observing that, although models generate longer explanations than humans, they demonstrate substantially less stylistic diversity. Our findings suggest that, while LLMs may serve as useful tools for generating diverse annotations, they should not be viewed as drop-in replacements for human annotators — particularly in applications requiring authentic representation of diversity in human judgments, such as NLI.

1 Introduction

A prominent line of NLP research examines the potential of large language models (LLMs) for annotation, either autonomously or in a collaborative setting with humans (Tan et al., 2024). Numerous LLM-based dataset generation pipelines have thus been introduced, incorporating prompt refinement and verification techniques to generate synthetic versions of existing datasets (Huang et al., 2025; Ye et al., 2022a,b). However, while such approaches appear promising in data augmentation scenarios,

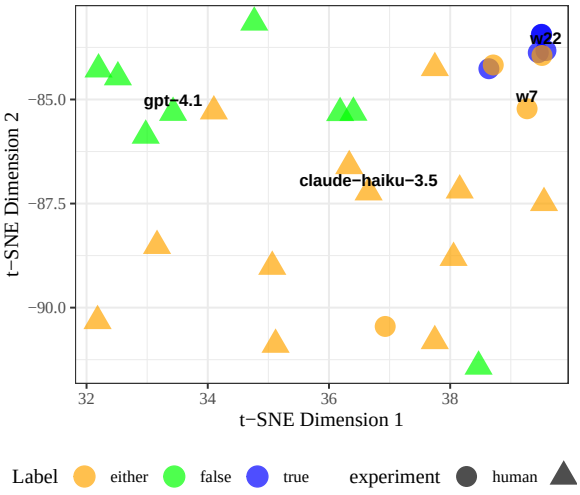

Figure 1: t-SNE visualization of e5 explanations embeddings for LIVENLI item `108083c`. While a majority of annotators agree in labeling the item as *either*, several models and humans diverge as to whether it can be exclusively *true* or *false*.

they generally lack the lexical diversity exhibited by human annotators and often fail to exceed the test-set performance of models trained on original, human-written training sets (Yu et al., 2023). Model-generated explanations, too, appear to offer limited utility: often increasing annotation time, cognitive load, and model reliance in LLM-assisted (re-)annotation efforts (Wang et al., 2024c), while failing to introduce novel information or support associated labels (Wiegreffe et al., 2022).

This line of research has demystified LLMs as all-purpose, expert annotators, but has primarily focused on the accuracy discrepancy between LLM and human-annotated data. One particularly overlooked aspect of the annotation process heretofore has been *human label variation* (HLV), wherein humans offer divergent judgments on a given item (Uma et al., 2021; Plank, 2022). Though, historically, such disagreements are often resolved by means of majority-voting (or other similar mechanisms), numerous studies have found that incorpo-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16526–16543

rating such information can aid in understanding models' predictive biases. To this end, Natural Language Inference (NLI) has proved to be an invaluable test-bed for understanding the effects of HLV, with studies reporting that models generally fail to capture annotator disagreement (Pavlick and Kwiatkowski, 2019), which often leads to a significant performance drop on high-variation items (Nie et al., 2020). In the era of LLM-assisted annotation, it thus becomes crucial to interrogate whether LLMs bear distributional similarity to human label variation — or, alternatively, whether they exhibit artificial consensus that obscures the ambiguity inherent in many annotation tasks.

In this study, we investigate how label variation manifests within *and* across LLMs for the NLI task, and how well such variation compares with what is observed for various human annotator groups. Though previous work has largely explored variation alignment by means of model-specific logits (Lee et al., 2023; Chen et al., 2024, 2025), we instead treat multiple model outputs — generated through various ensembling strategies — as distinct annotators (see Figure 1 for an example). This approach enables us to examine which ensemble-based sampling methods best compare to the inherent ambiguity and disagreement present in human judgments, as well as the extent to which this is reflected in explanations. We formalize this line of inquiry with the following research questions:

1. Which ensemble-based sampling strategy best reflects the distribution and diversity of human label variation?
2. How does the distribution of label variation in LLMs compare to that of human annotators?
3. To what extent does label variation in LLMs correspond to variation in their explanations?

In studying these research questions on English NLI data, we hope to determine whether — and how — LLM ensembles can serve as viable proxies for diverse human annotator groups. We make all of the model-generated responses available for download as a new dataset called `PartyNLI`.[1]

2 Related work

LLMs as annotators Numerous methods for leveraging LLMs for annotation tasks have emerged in recent years, with many showing initial promise (Ye et al., 2022b,a; Meng et al., 2022). For example, Huang et al. (2025) propose a LLM-oriented data generation pipeline that employs various self-refinement and evaluation steps to produce higher quality data. However, while such approaches are generally effective for math-oriented or coding tasks, it has been shown that LLMs typically fall short when it comes to producing diverse, creative output that is characteristic of human annotators (Yu et al., 2023; Giulianelli et al., 2023). Baumann et al. (2025) demonstrate that different model configurations can yield potentially conflicting task solutions, affecting the trustworthiness of their decision-making processes.

To account for these potential shortcomings, another line of research has explored the potential of human-LLM collaboration in annotation. For example, Liu et al. (2022) select ambiguous NLI items and instruct a model to generate more samples according to the reasoning patterns present therein, leading to better out-of-domain performance. Wang et al. (2024c) propose a more generalized, multi-step framework in which LLM annotations are passed to a verifier system, the output of which is used to identify items for human re-annotation. They also show that model-generated explanations tend to be occasionally detrimental to annotation efforts, leading humans to select incorrect labels and become cognitively overloaded, echoing earlier findings by Bansal et al. (2021).

Label variation in NLI Plank (2022) argues that the concept of human label variation is a bottleneck for progress in NLP, as traditional methods tend to optimize for a single *ground truth* and inevitably discount notions of ambiguity or genuine disagreement across annotators. Pavlick and Kwiatkowski (2019) demonstrate that this is a persistent issue for a variety of English NLI datasets, where systems dramatically fail to learn human-like models of uncertainty. Nie et al. (2020) attempt to corroborate this on a larger scale, collecting 100 new human judgments per item for 3 different English NLI datasets. Ultimately, they show that, while models vary in the alignment of their softmax layer to the distribution of human judgments, all models perform exceptionally poorly on high-entropy data points. Building on these findings, recent work has taxonomized NLI variation: Jiang and Marneffe (2022) introduce a 10-point disagreement taxonomy utilized in LIVENLI (Jiang et al., 2023a); Weber-Genzel et al. (2024) propose VARIERR, which disentangles annotation error from dis-

[1] https://huggingface.co/datasets/aseaofcars/PartyNLI

agreement; Hong et al. (2025) classify instances of *within-label* disagreement.

3 Experimental design

We investigate the extent to which LLM label variation compares to HLV in NLI. Previous work has sought to accomplish this by means of extracting and normalizing model logits corresponding to the NLI label space (*entailment, neutral, contradiction*) and comparing the resulting (probability) distribution with the distribution of labels assigned by a group of annotators (Lee et al., 2023; Chen et al., 2024). However, this method of distribution expression has since been found to be poorly calibrated (Meister et al., 2025; Xia et al., 2025) and generally inconsistent across prompting settings (Wang et al., 2024b). In this study, we opt instead to treat models as *participants* prompted the same way as the humans — allowing us to directly assess the distribution of their judgments. After obtaining a model-generated dataset, we compare human and model label distributions (and their associated explanations) by means of various distributional metrics. We divide our experiments into three categories: 1) the predictions of a SINGLE model across 5 prompting runs; 2) a MODELMIX, wherein we sample generations from a combination of LLMs; and 3) a parameter setting, wherein we vary generation parameters (temperature (τ) and model size).

Data We use LIVENLI, a collection of 122 re-annotated MNLI items chosen to encompass a range of disagreement in the original dataset (Jiang et al., 2023a). For each item (a premise and hypothesis pair), crowd-sourced annotators were asked whether the hypothesis is "most likely to be true/-false/either true or false" (according to the standard NLI labeling scheme of "entailment", "contradiction", or "neutral"), with the option to choose multiple labels. They were instructed to write a free-text explanation for every label they chose, as well as to highlight the words in the premise and/or hypothesis that are relevant to their explanations. Figure 5 in Appendix A shows the exact instructions provided to annotators.

Each LIVENLI item features 10 distinct annotations, making it a suitable choice for studying label variation. To account for variation across different annotator populations, we also consider label distributions of the 50 LIVENLI items that come both from CHAOSNLI (Nie et al., 2020) (100 annotators) and the original MNLI dataset (5 initial annotators) (Williams et al., 2018). Although LIVENLI allows for the selection of multiple labels per item, both CHAOSNLI and MNLI instruct annotators to provide a single label. To account for this, we convert each multi-label response to a single label by choosing *true* or *false* if these labels were provided in isolation, and *either* otherwise.

Zero-shot prompting To simulate the human annotation process, we perform zero-shot prompting on all model variants outlined above. We avoid few-shot or chain-of-thought (CoT) prompting here, as we are interested in exploring LLM label and explanation variation out-of-the-box, and because doing so would condition the model response on information that is extraneous to the original annotation instructions. For input, models are provided a `user` prompt, which contains instructions extracted from the LIVENLI data collection interface. We also pass a `system` prompt, which describes the model's role for the task, as well as a set of formatting specifications. For the MODELMIX and model size experiments, we run five trials across all models, amounting to $122 \cdot 5 = 610$ annotations per model. Our prompts are available in Appendix B.

Model selection To maximize sample diversity, we work with a wide range of open and closed-source LLMs. We select the 10 highest-ranked models[2] from the LMArena leaderboard,[3] across different providers — one per lab to avoid pre/post-training biases. We refer to this sample as MODELMIX$_{\text{hi}}$. Additionally, we complement this sample with another selection of 10 older,[4] smaller, and less-performant models, choosing the previous version of MODELMIX$_{\text{hi}}$ models where possible (e.g. `claude-haiku-3.5` as a complement to `claude-opus-4`). In the case of open source models, we opt to work with their smallest corresponding version within the 8B-20B size range. This sample is called MODELMIX$_{\text{lo}}$, and the combination of all 20 models is MODELMIX$_{\text{all}}$. Here, it is important to note that differences in model design are not directly comparable to human demographic diversity; we therefore make no claims that MODELMIX$_{\text{all}}$ constitutes a proxy for the human annotator population.

For the model size experiments, we select LLM families for which at least four size variants

[2] As of July, 2025

[3] `https://lmarena.ai/`

[4] Released before December 2024.

are available: `gemma3`, `deepseek-r1`, and `qwen3`, which come in 4, 6, and 7 sizes, respectively.[5] Here, we hypothesize that model scaling bears an effect on how models classify NLI items, and, by extension, that this manifests as label variation across sizes. In terms of the temperature modulation experiments, we choose the top and bottom models ranked by label distribution entropy (from both MODELMIX$_{\text{hi}}$ and $_{\text{lo}}$) as representatives of varying degrees of labeling consistency. This selection comprises of `grok-3-mini`, `qwen3`, `gemini-2.0-flash`, and `kimi-k2`. Table 1 gives our full model selection.

MODELMIX$_{hi}$	MODELMIX$_{lo}$
`claude-opus-4`	`claude-haiku-3.5`
`command-a`	`command-r`
`gemini-2.5-pro`	`gemini-2.0-flash`
`gpt-4.1`	`gpt-4o-mini`
`grok-4`	`grok-3-mini`
`llama-4-maverick`	`llama-3.1-8b`
`mistral-medium`	`ministral-8b`
`qwen3-235b`	`qwen2.5-14b`
`kimi-k2`	`olmo2-13b`
`deepseek-r1-671b`	`phi4-14b`
Model	Size (B)
`deepseek-r1`	7, 8, 14, 32, 70, 671
`gemma3`	1, 4, 12, 27
`qwen3`	0.6, 1, 4, 8, 14, 32, 235

Table 1: *Top*: MODELMIX models. *Bottom*: size experiment models and sizes (in billions of parameters)

Generation parameters When making API calls to model providers, we do not pass additional parameters besides the system and user prompts, except in the temperature modulation experiments where we provide a range of 10 distinct τ values to match the number of human annotations per item. It is important to note that temperature is theoretically bounded on $[0.0, \infty)$, though values above 2.0 often tend to result in nonsensical output (Holtzman et al., 2020), and are therefore clamped to a reasonable upper bound $\tau_{\max}$ (often 1.0 or 2.0) by providers. To account for this, we generate 10 equally spaced values within the range $[0.05 \cdot \tau_{\max}, 0.95 \cdot \tau_{\max}]$, allowing us to generate a dynamic range of τ values that respects model-specific calibration.[6]

[5]We omit the smallest `deepseek-r1` version (1.5B), as it failed at outputting responses in a structured `JSON` format.

[6]While the `grok` API accepts $[0, 2]$ as a range for τ, we find that the model does not generate coherent output beyond $\tau = 1.5$, so we set this as $\tau_{\max}$.

4 Measuring label variation

We use entropy (H) as a measure of label variation within each experimental setting. We also report Krippendorff's α for all model settings, as well as for the LIVENLI annotators, for which we have access to all 48 annotator identifiers. For distribution *alignment*, we calculate the JS divergence between model and human label distributions, where values closer to 0 indicate closer alignment. Since all of these metrics (except α) are computed at the *item* level, we report their mean and standard deviation over all items. Additionally, we use the Kendall correlation coefficient (Cor.) to assess whether items that are difficult for humans are similarly difficult for models, comparing per-item label distribution entropies across both groups. For the SINGLE setting, we report statistics for the model with the best JS score with respect to humans (`grok-3-mini`), with all remaining model scores shown in Appendix C.

4.1 Results

Table 2 displays the label variation results across all experimental settings for the 50 items common to LIVENLI, CHAOSNLI and MNLI. SINGLE (`grok-3-mini`) does not demonstrate high similarity to any human annotator population, yielding low *within*-model variation (low H, high α) as well as high divergence with human judgments (high JS). Thus, when only considering evaluation across identically repeated trials, single models do not exhibit similar label diversity as compared with that of humans.

Temperature modulation We observe a slight change in `grok-3-mini`'s behavior when varying τ. H increases with respect to SINGLE, indicating that more variation is introduced across runs when larger τ values are considered. We observe a similar pattern for `qwen3`, whose responses likewise appear to diversify when modulating generation parameters. Conversely, both `gemini-2.0-flash` and `kimi-k2` appear to be considerably well-calibrated to changes in τ, exhibiting very little added variation when compared to sampling with default parameters (see Appendix C for details).

Model size All three model families in the SIZE experiment exhibit higher H and lower JS than the τ settings. Notably, α is likewise much lower here, indicating large disagreement among model

Experiment	Parameters	LiveNLI H	LiveNLI Cor. ↑	LiveNLI JS ↓	LiveNLI α	ChaosNLI H	ChaosNLI Cor. ↑	ChaosNLI JS ↓	MNLI H	MNLI Cor. ↑	MNLI JS ↓
Human	–	0.64_{23}	–	–	0.26	0.72_{13}	–	–	0.72_{10}	–	–
SINGLE	–	0.19_{28}	0.00	0.33_{15}	0.67	0.19_{28}	−0.01	0.34_{16}	0.19_{28}	0.26	0.37_{14}
MODELMIX$_{\text{hi}}$	–	0.52_{23}	0.07	0.24_{14}	0.38	0.52_{23}	0.07	0.24_{13}	0.52_{23}	0.14	0.28_{14}
MODELMIX$_{\text{lo}}$	–	0.60_{18}	0.03	0.24_{14}	0.19	0.60_{18}	−0.15	0.24_{11}	0.60_{18}	0.07	0.22_{15}
MODELMIX$_{\text{all}}$	–	0.63_{15}	0.04	0.21_{13}	0.27	0.63_{15}	−0.01	0.21_{10}	0.63_{15}	0.19	0.23_{13}
`gemini-2.0`	$\tau_{\max} = 2.0$	0.07_{21}	0.07	0.41_{20}	0.92	0.07_{21}	0.08	0.39_{14}	0.07_{21}	0.11	0.45_{13}
`grok-3-mini`	$\tau_{\max} = 1.5$	0.25_{26}	0.08	0.33_{14}	0.70	0.25_{26}	−0.06	0.34_{15}	0.25_{26}	0.36	0.38_{13}
`kimi-k2`	$\tau_{\max} = 1.0$	0.12_{18}	−0.04	0.35_{17}	0.87	0.12_{18}	0.19	0.40_{13}	0.12_{18}	0.10	0.44_{10}
`qwen3`	$\tau_{\max} = 1.0$	0.28_{28}	0.00	0.33_{17}	0.60	0.28_{28}	−0.15	0.37_{15}	0.28_{28}	0.05	0.38_{14}
`deepseek-r1`	size	0.49_{28}	−0.03	0.26_{15}	0.30	0.49_{28}	−0.08	0.28_{13}	0.49_{28}	0.11	0.29_{16}
`gemma3`	size	0.49_{25}	0.20	0.28_{17}	0.11	0.49_{25}	−0.16	0.27_{15}	0.49_{25}	0.02	0.26_{16}
`qwen3`	size	0.62_{21}	−0.08	0.25_{14}	0.25	0.62_{21}	0.18	0.23_{11}	0.62_{21}	0.12	0.25_{15}

Table 2: Aggregate results across all experimental settings over the 50 items common to LIVENLI, CHAOSNLI and MNLI (MODELMIX and size metrics are averaged across trials). Highlighted cells correspond to model settings that yield highest similarity to human metrics. Subscripted values denote average standard deviation per trial in units of the least significant digit — e.g. 23 is equal to 0.23.

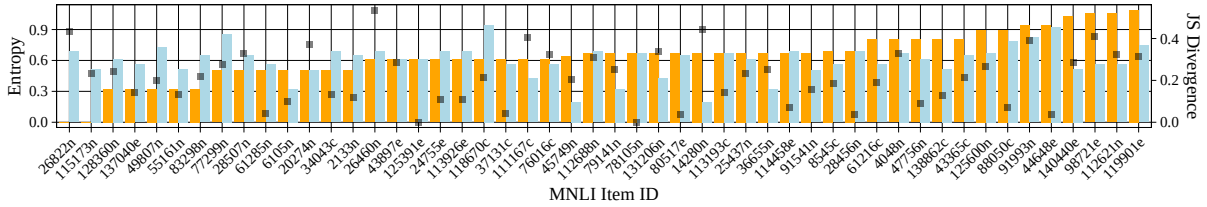

Figure 2: Entropy per-item for human and MODELMIX$_{\text{all}}$ models on the 50 items common to LIVENLI, CHAOSNLI and MNLI. Jensen-Shannon divergence is overlaid on top of corresponding items with ■.

variants. This is expected, given that each response originates from a distinct model (rather than different sampling parameters), enabling its unique architectural characteristics and training procedure to influence the resulting label distributions. Indeed, these models differ substantially in training tokens (e.g., 13 trillion between `gemma3` sizes), post-training procedures, and architectures (Team et al., 2025; Guo et al., 2025; Yang et al., 2025). Ultimately, `qwen3` demonstrates not only the lowest JS, but also the largest H, likely due to offering the most models, as well as the largest spread of model sizes.

MODELMIX Perhaps predictably, the MODELMIX experiments yield the most diverse judgments, resembling what is observed across human annotator groups. This setting accounts for *participant diversity*, wherein models are purposely sampled from a large variety of LLM providers. As a result, MODELMIX is capable of capturing a broad range of judgment patterns and biases across different model architectures, training methodologies, and developer philosophies. While MODELMIX$_{\text{hi}}$ and MODELMIX$_{\text{lo}}$ exhibit slightly different behaviors, the combination of all 20 models generally yields the closest overall distributional match to humans (in terms of H, JS, and α). However, the most notable exception here, and nearly all other model settings, is Cor., which is strikingly low. This indicates that, while distributional similarity for labels varies across settings (JS), models largely do not tend to agree with humans in their classification of difficult or ambiguous items. Figure 2 illustrates this phenomenon for LIVENLI, showing that human and model entropies rarely align, and that low JS values can be observed despite large differences in H.

4.2 Analysis of MODELMIX Responses

We now conduct an analysis of the labeling patterns of MODELMIX$_{\text{all}}$ over the entire LIVENLI dataset. Figure 3 (left) depicts the overall distribution of labels per model and six human participants who annotated more than 100 (out of the 122) LIVENLI items (left-top). We notice a drastic difference between the human and model label distributions: the former is more balanced across labels and comparable across participants, while the latter is highly skewed towards the *either* label — comprising the majority of most models' answers. Specifically, `ministral-8b`, `gpt-4o-mini`, and `llama-3.1-8b`

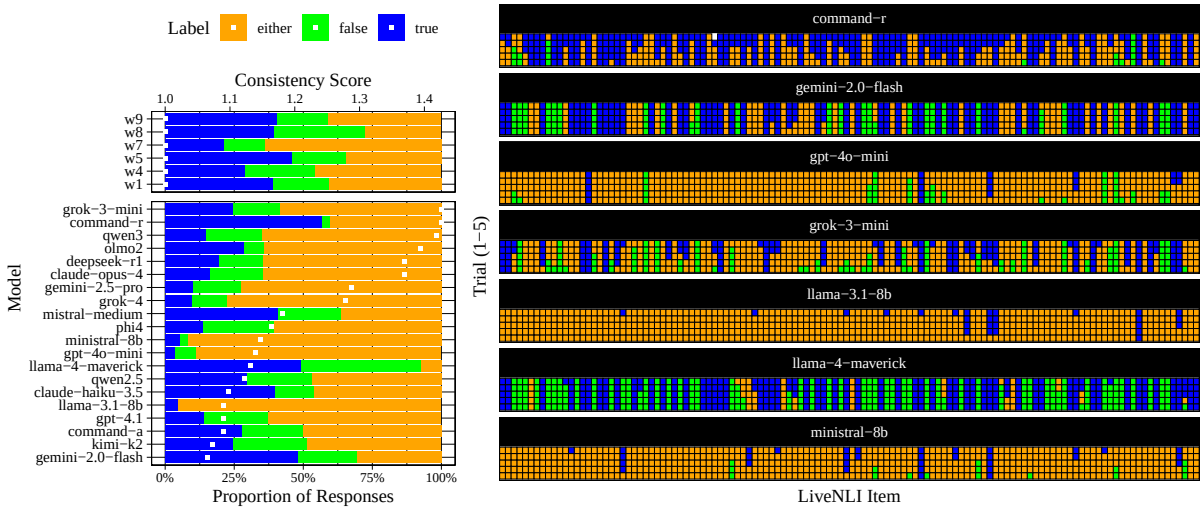

Figure 3: *Left*: Overall distribution of labels by model. Models are ordered by average response consistency (least consistent on top), with overlaid white squares □ representing average consistency values; *Right*: Model classification across trials (each cell represents a single classification within a trial).

all select *either* for more than 95% of the entire LIVENLI dataset, with the latter failing to classify a single item as *false*. This suggests that models often attempt to justify choosing both *true* and *false*, even in cases where only a single analysis is valid. Interestingly, this behavior is not uniform across all models: for example, `llama-4-maverick` selects only *true* or *false* for over 95% of its answers, indicating a bias towards decisiveness. While it is impossible to diagnose the exact reason for a given model's behavior, we suspect that it is highly dependent on the post-training pipeline.

We also notice large differences in *consistency* between models, defined as the average amount of unique labels across trials. In our case, a score of 1 indicates perfect consistency across trials, with models deterministically assigning the exact same label to each item. On the other hand, a consistency of 3 implies that a model covers the full range of possible labels, suggesting that it might assign labels at random. With this in mind, `gemini-2.0-flash` generally proves to fare the best (1.07), offering perfectly consistent judgments for 94% of items, while covering a balanced distribution of labels in a manner similar to that of humans. The same can be said for `kimi-k2` (1.07), though it exhibits larger overall skew towards *either*. Conversely, `grok-3-mini` is the *least* consistent model (1.43), generally exhibiting overall confusion between *true*/*false* and *either*. The same can be said for `command-r` (1.43), though the label space here is even further restricted, with the model seemingly only deciding between *true* and *either*.

5 Measuring explanation variation

Unlike labels, variation in model-generated explanations is not as easily quantifiable or interpretable. Following previous studies, we employ Jaccard similarity between word unigrams at the item level (JAC) to account for surface-level variation in lexical choice. We note, however, that many model-generated explanations quote text from the premise and hypothesis, which can potentially inflate pairwise similarity. We thus *de-lexicalize* explanations with respect to their corresponding LIVENLI item, removing trigrams that overlap directly with the premise/hypothesis. In addition to JAC, we also report word count (w_c) and vocabulary size (V) (both delexicalized). While these measures are limited to lexical properties and do not reflect deeper semantic variation, they nonetheless allow us to compare how explanations are expressed across both annotator populations.

Beyond lexical overlap, a broader issue affecting the measurement of explanation variation is topic bias. This is particularly salient when working with general-purpose encoders, as inter-item explanations generally yield much higher similarity to each other than to any other unrelated item, even if their associated labels differ. Indeed, we find that different general-purpose embeddings tend to cluster explanations perfectly by item (see Appendix G for details). We thus adopt three distinct vector-based representations for the measurement of explanation variability:

1. EMB_{e5}: general-purpose text embeddings

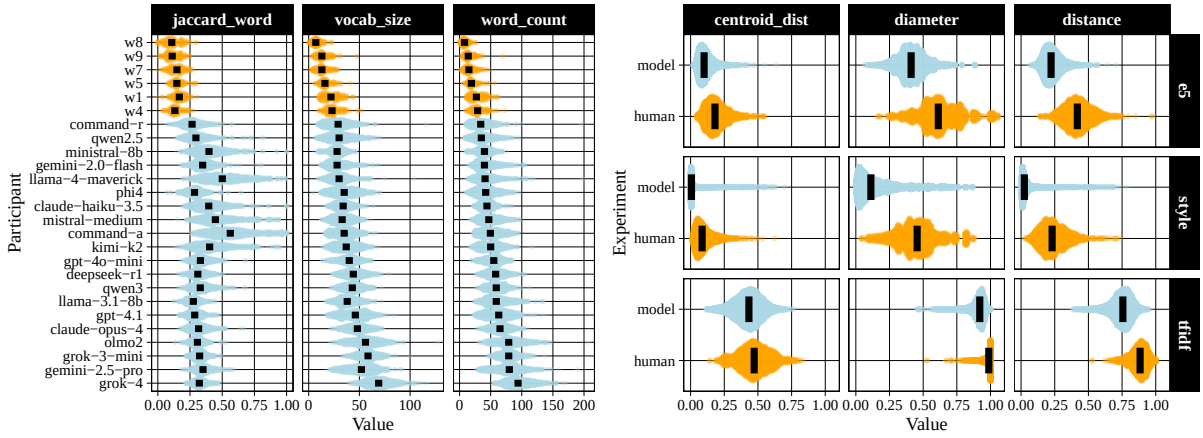

Figure 4: *Left*: Distribution of lexical metrics across LIVENLI, for some human annotators ($n > 100$ items) and MODELMIX$_{\text{all}}$ models. *Right*: Distribution of vector-based metrics within human and MODELMIX$_{\text{all}}$ groups.

from Wang et al. (2024a), from which we subtract the embedding for the corresponding (concatenated) premise + hypothesis string.

2. EMB$_{\text{style}}$: embeddings from Patel et al. (2025), capturing stylistic differences across authors, while minimizing topic similarity.
3. EMB$_{\text{tfidf}}$: sparse TF-IDF vectors fit on de-lexicalized explanations (as done for JAC).

As a basic measure, we report average pairwise cosine distance (D_{pair}), allowing us to verify the extent to which explanations differ across annotators. Additionally, we calculate the average distance to (explanation) centroid (D_{cent}), which provides further insight into the overall coherence or diversity of the explanation set. Lastly, we report the maximum pairwise distance (D_{max}), (otherwise known as diameter in graph theory), which identifies the most divergent pair of explanations and provides a measure of the full range of variability. Importantly, our calculations are made at the label level for each item, so as to avoid conflating disagreement with variation. We report results for human participants and MODELMIX$_{\text{all}}$ over LIVENLI (all remaining results can be found in Appendix H).

5.1 Results

Figure 4 (left) displays the results for explanation variation measures. Visualizing the distribution of lexical metrics across humans and models, there is a clear distinction between the two groups. Every model exhibits higher median w_c, V, and JAC scores across LIVENLI items, indicating that models generate longer, more lexically-rich explanations, even after de-lexicalizing with respect to the items. Interestingly, we observe a range of explanation lengths and vocabulary sizes within both groups: participant `w9` returns a median 14 tokens per explanation, while `w4` produces more than double that amount: 30. The range for models is even more stark, with `grok-4` returning the highest median word count of any model (79), as well as the longest overall explanation (137).

The embedding-based metrics (Figure 4 right) show another consistent trend: models in MODELMIX$_{\text{all}}$ demonstrate far less *variation* across all metrics and embedding types when compared to humans. This is particularly interesting given the above findings, which show that models tend to generate far longer and more lexically diverse explanations. For example, the average distance between EMB$_{\text{e5}}$ embeddings is significantly smaller for models than humans, suggesting that all 20 models may rely on similar descriptive strategies for explanations. This is further corroborated by D_{cent} demonstrating that model explanations tend to cluster more tightly, and D_{max} which indicates smaller overall spread and fewer outliers.

Interestingly, while the magnitude of average distance remains largely comparable across EMB$_{\text{e5}}$ and EMB$_{\text{style}}$, the overall distribution of values proves to be strikingly different. Specifically, the median average distance across models is 0.02 in the latter case, suggesting that models adopt a singular authorship "style", while human explanations are more individually distinct (0.23). This insight is backed up by D_{max}, which follows a similar, left-skewed distribution with a low median value of 0.11 (compared to 0.46). Finally, although the difference in variability between models and humans is the most subtle for EMB$_{\text{tfidf}}$, we nonetheless ob-

Premise **Hypothesis**	Do you want to see historic sights and tour museums and art galleries? You wouldn't like to visit historic places, museums, and art galleries, would you?
w22 (true)	The heart of both of these questions seems to match.
w7 (either)	The context is asking if someone would like to do these things, and there is no indication of whether or not they do.
gpt-4.1 (false)	The context asks if you want to see historic sights and tour museums and art galleries, which presupposes an interest or desire. The statement, however, asserts (in negative form) that you would not like to visit historic places, museums, and art galleries. Since the context implies a positive preference and the statement implies a negative one, assuming the context is true, the statement is most likely false.
claude-haiku-3.5 (either)	The statement uses a negative question format which creates ambiguity. The context suggests an interest in historic sites and museums, while the statement uses a double-negative phrasing ('wouldn't like') that could imply either agreement or disagreement with the original context. The interpretation depends on the tone and intention of the speaker, making it possible to read the statement as either confirming or contradicting the original interest.

Table 3: Selected participant explanations from Figure 1.

serve the same overall trend. The distances here are generally larger, given that tfidf vectors are sparse, and only account for lexical overlap. With this in mind, the very high values of $\text{EMB}_{\text{tfidf}}$ returned for humans remains largely consistent with the story so far, indicating very low lexical overlap.

5.2 Qualitative analysis of explanation content

	Reason. Type			Human Label			Model Label		
	Sh	Hu	Mo	N	C	E	N	C	E
78105n	3	0	1	0.4	0.6	0.0	0.4	0.6	0.0
80517e	1	2	1	0.6	0.0	0.4	0.65	0.0	0.35
61285n	3	0	1	0.8	0.0	0.2	0.75	0.0	0.25
73518e	4	0	1	0.6	0.3	0.1	0.4	0.35	0.25
49807n	3	0	2	0.9	0.0	0.1	0.75	0.1	0.15
126837e	1	1	1	0.0	0.0	1.0	0.15	0.0	0.85
140704e	3	1	0	0.6	0.3	0.1	0.9	0.05	0.05
76016c	2	1	0	0.3	0.7	0.0	0.75	0.25	0.0
20274n	2	0	0	0.8	0.2	0.0	0.8	0.0	0.2
101940n	3	1	3	0.6	0.0	0.4	0.42	0.42	0.16
26460n	3	1	5	0.3	0.0	0.7	0.75	0.2	0.05
108083c	1	1	5	0.5	0.0	0.5	0.6	0.4	0.0

Table 4: *Left*: Explanation types shared (Sh) by both humans and models, only put forth by humans (Hu) or models (Mo) for 12 LIVENLI items. *Center, Right*: Distribution of neutral (N), contradiction (C), and entailment (E) labels by Humans and Models.

We further explore whether LLMs generate explanations with reasoning patterns comparable to those of humans. We selected 12 LIVENLI items (10% of the data) for manual analysis: six items with the highest and lowest three JS scores, and six sampled uniformly at random (accounting for all humans and the first trial of $\text{MODELMIX}_{\text{all}}$). Table 4 displays the counts of reasoning types shared between humans and models, as well as novel ones employed by each annotator group. We find that, for most items, LLMs come up with similar reasoning types to those of humans, though they may not necessarily assign the same label (green) — e.g., a subset of models and humans diverge in classifying 20274n as *true* or *false*, but do not exhibit any novel explanations. We also observe the opposite phenomenon, where both annotator groups demonstrate almost perfect JS alignment, but nonetheless introduce novel reasoning types (yellow). Indeed, the low JS items (red) demonstrate the starkest contrast between model and human explanations, with both groups exhibiting non-overlapping bimodal label distributions. In such cases, models appear to introduce many more novel reasoning types, though these tend to relate to concepts humans don't generally mention, like tone.

For illustrative purposes, we display four explanations for the lowest JS item 108083c in Table 3, which complements Figure 1. Interestingly, w22's explanation matches semanticists' analysis of the entailment relationship between questions (Groenendijk and Stokhof, 1984). We observe that claude-haiku-3.5 and w7 agree on *either*, and that the gist of the explanations is similar: the answer to the question is not known. However, claude-haiku-3.5 also produces a hallucination (*statement uses a double-negative*), and discusses the speaker's tone, which no humans mention. gpt-4.1 justifies its *false* prediction by interpreting the hypothesis as asserting that the addressee does not want to visit historic places. However, while the hypothesis implies a dispreference of the addressee for visiting historic places, it certainly does not assert it, as the hypothesis is a question. In fact, negative tag questions like this are commonly used to make playful offers — for instance, asking a child who loves ice cream, *You wouldn't want some ice cream, would you?*

6 Conclusion

In this study, we investigated label and explanation variation in the context of LLM-based annotation. Rather than attempt to directly align LLMs with human label distributions, we analyzed various settings in which LLM judgments could be solicited, focusing on variation in generation parameters, within-family model size differences, and combinations of distinct LLMs. We found that the extent of label variation depends largely on the

choice of individual models, with some demonstrating much higher labeling consistency than others. Furthermore, we showed that increased variation can be observed when eliciting judgments across models within the same family, or across ensembles of distinct models. Crucially, however, while various distributional metrics imply that model and human variation is characteristically similar, we demonstrated that models tend to disagree in their judgment of difficult items and exhibit striking idiosyncrasies that are not observed in humans.

Given the attested diversity of labeling tendencies, we conducted another experiment, wherein we investigated the extent of *explanation* variation across models. Interestingly, we found that, while models appear to generate longer explanations than humans, they exhibit far less variation in lexical choice and writing style. In our qualitative analysis of a small sample of items, we observed that, while models often align with human explanation contents, they can be more imaginative, but also sometimes wrong. Such findings open the door for future research on explanation evaluation in the face of ambiguity, disagreement, or hallucinations.

7 Limitations

There are several limitations to our work. Chief among these is that we report our results with respect to a single dataset (LIVENLI). However, though we surveyed multiple NLI datasets for our experiments, LIVENLI was the only one that covered every one of our prerequisites, such as offering an opportunity to study label *and* explanation variation. Nonetheless, we acknowledge that this limited sample may contain biases or artefacts of which we were not immediately aware. By extension of this, we likewise only study one task (NLI), which, despite being well studied, features many inherent problems (Pavlick and Kwiatkowski, 2019). For future work, we would like to extend our method towards other, potentially more complicated tasks, to verify if our results hold. One potential candidate is CoS-E (Rajani et al., 2019), which collects explanations for extractive QA.

Regarding the methodology described in Section 5, we acknowledge that our experiments target variation on a largely lexical and form-oriented level. While our results are informative, they are nonetheless incapable of capturing disparate lines of reasoning and argumentation expressed in free-text explanations. This is certainly desirable, but is difficult to capture by automated means and aggregate metrics (e.g. BLEU, ROUGE, METEOR). Future work will explore this avenue further, assessing the feasibility of questions-under-discussion (QUD) as a means of eliciting "reasoning types" in explanations (Namuduri et al., 2025).

Lastly, we acknowledge a common limitation of NLP research that also applies to our study: working with monolingual English data. While similar degrees of human label variation have been reported for Chinese (Hu et al., 2020) and Japanese (Yanaka and Mineshima, 2022) NLI annotation efforts, it is entirely possible that the results described here would not be fully robust against a thorough multilingual evaluation. To this end, it has been shown that annotators make different inferential judgments based on cultural background (Huang and Yang, 2023). We leave this as an additional avenue for future work.

8 Acknowledgements

This work was supported by the Fonds de la Recherche Scientifique – FNRS under Grant n° F.4511.25. Marie-Catherine de Marneffe is a Research Associate of the Fonds de la Recherche Scientifique – FNRS.

References

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation. *arXiv preprint*. ArXiv:2509.08825 [cs].

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10777–10802, Vienna, Austria. Association for Computational Linguistics.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "Seeing the Big through the Small": Can LLMs Approximate Human Judgment Distributions on NLI

from a Few Explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *arXiv preprint*. ArXiv:1904.09751 [cs].

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2025. LiTEx: A linguistic taxonomy of explanations for understanding within-label variation in natural language inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34065–34085, Suzhou, China. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, and Lichao Sun. 2025. DataGen: Unified Synthetic Dataset Generation via Large Language Models. *arXiv preprint*. ArXiv:2406.18966 [cs].

Nan-Jiang Jiang and Marie-Catherine De Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine De Marneffe. 2023a. Ecologically Valid Explanations for Label Variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023b. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 462–477, Red Hook, NY, USA. Curran Associates Inc.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. QUDsim: Quantifying Discourse Similarities in LLM-Generated Text. *arXiv preprint*. ArXiv:2504.09373 [cs].

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*. ArXiv:2402.05672 [cs].

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024c. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, Honolulu HI USA. ACM.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating Annotation Error from Human Label Variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Yuxi Xia, Pedro Henrique Luz De Araujo, Klim Zaporojets, and Benjamin Roth. 2025. Influences on LLM calibration: A study of response agreement, loss functions, and prompt styles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3740–3761, Vienna, Austria. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. ProGen: Progressive Zero-shot Dataset Generation via In-context

Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 55734–55784, Red Hook, NY, USA. Curran Associates Inc.

A LIVENLI annotation instructions

Figure 5 depicts the LIVENLI interface, which was converted to a plain-text `user` prompt and passed to all models.

B Prompt Templates

Figures 6 and 7 show the exact prompts used for the experiments in this study.

C Label Variation per Model

Table 5 shows the results for all label variation metrics across individual models. While we make note of variation across models, no individual model matches the performance of the ensemble experiments outlined in Section 4.1.

D RANDOMMIX

Though it might appear that collecting the largest possible sample of models is the best way to ensure variation in annotation tasks, we acknowledge that this is often not practical due to high API inference costs. As such, we run an additional experiment wherein we simulate 1000 random draws of 10 $\text{MODELMIX}_{\text{all}}$ models in order to verify the stability of random samples against our metrics (an experiment to which we refer to as RANDOMMIX). As Figure 8 shows, each metric lies largely in the same range as either MODELMIX across 1000 trials, with JS showing the least variability. These findings indicate that a sample of distinct LLMs can provide an acceptable proxy of human label variation.

E Label Distributions (τ)

Figure 9 depicts the per-item consistency for all models in the τ parameter experiments (we omit top-p as it is very similar to this plot). This corroborates our main findings, showing that `grok-3-mini` and `qwen3` are largely inconsistent across varying values of τ, while the two other models are generally stable.

F Label Distributions (SIZE)

Figure 10 depicts the per-item consistency for all models in the $size$ experiments. Interesting, we find that some models exhibit the same sorts of biases discussed in Section 4.2. For example, `qwen3-0.6B` largely answers decisively, yet inconsistently (per our metric), exhibiting a clear prefer-

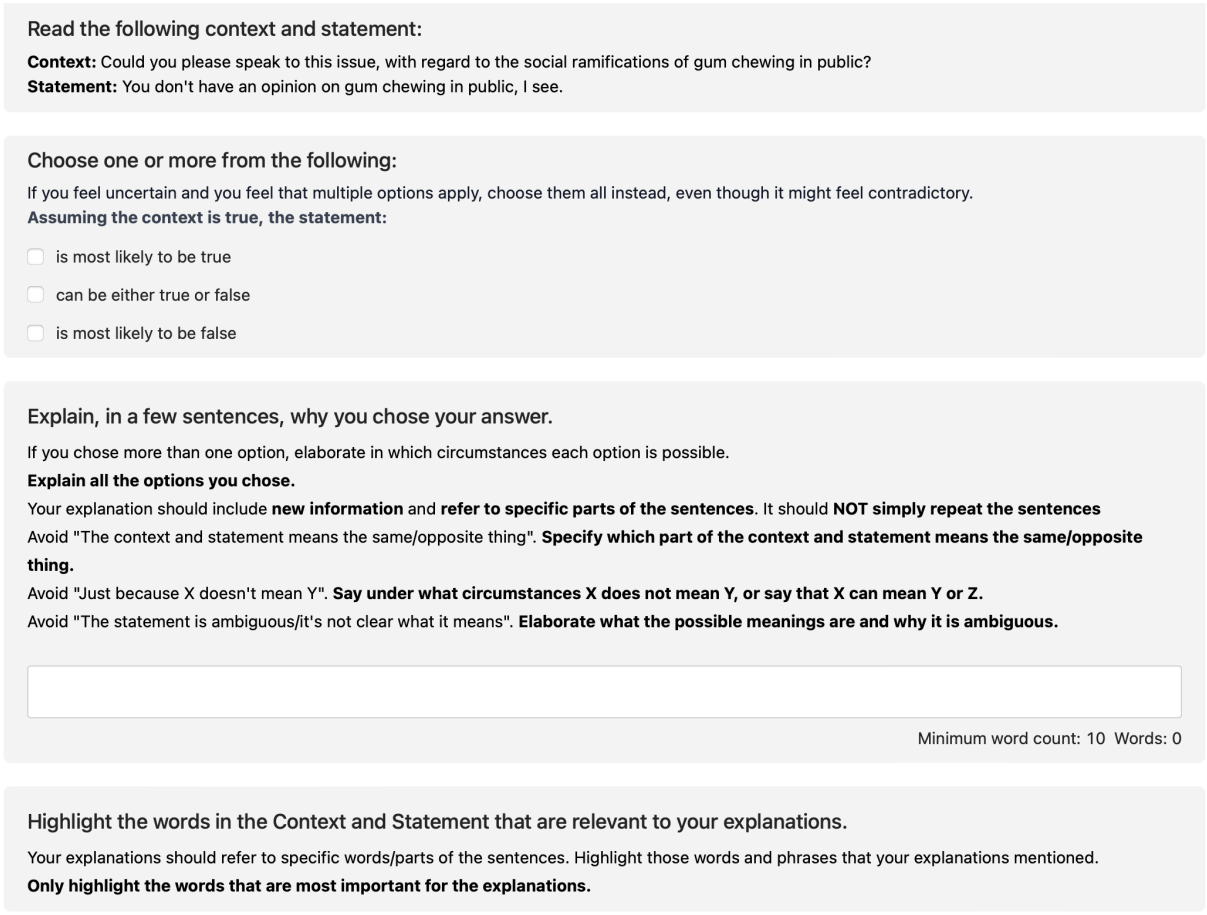

Figure 5: Screenshot of the LIVENLI data collection interface. Taken from (Jiang et al., 2023b)

System Prompt

```
You are an expert at natural language inference tasks.
Always provide your responses as valid JSON that can be parsed by Python's json.loads() function.

Your response format should be:
{
    "classification": {
        "label": "...",
        "explanation": "..."
    },
    "highlights": {
        "context": "...",
        "statement": "..."
    }
}

For classification labels:
- If selecting multiple classes, sort them alphabetically and join with hyphens (e.g., "either-false-true")

For highlights:
- Mark highlighted segments with '<<' at the beginning and '>>' at the end
- Only highlight the most important words/phrases mentioned in explanations

CRITICAL: Respond with ONLY the raw JSON object. Do not use markdown formatting, code blocks, or any wrapper text.

Do NOT wrap your response in:
- ```json ... ```
- ``` ... ```
- Any markdown formatting
- Any explanatory text before or after the JSON
```

Figure 6: System prompt provided to all models.

ence *true* or *false*. Conversely, `gemma3-1B` shows the opposite trend in consistently answering *either*.

User Prompt

```
Read the following Context and Statement (introduced respectively by the [CONTEXT] and [STATEMENT] tags):
[CONTEXT] {}
[STATEMENT] {}

Choose one or more from the following:

If you feel uncertain and you feel that multiple options apply, choose them all instead, even though it might feel
contradictory.
Assuming the context is true, the statement:

- true: is most likely to be true
- either: can be either true or false
- false: is most likely false

Explain, in a few sentences, why you chose your answer.
If you chose more than one option, elaborate in wich circumstances each option is possible.
Explain all the options you chose.
Your explanation should include new information and refer to specific parts of the sentences. It should NOT simply
repeat the sentences.
Avoid "The context and statement means the same/opposite thing". Specify which part of the context and statement
means the same/opposite thing.
Avoid "Just because X doesn't mean Y". Say under what circumstances X does not mean Y, or say that X can mean Y or Z.
Avoid "The statement is ambiguous/it's not clear what it means". Elaborate what the possible meanings are and why it
is ambiguous.

Highlight the words in the Context and Statement that are relevant to your explanations.
Your explanations should refer to specific words/parts of sentences.
Highlight those words and phrases that your explanations mentioned.
Only highlight the words that are most important for the explanations.
```

Figure 7: User prompt provided to all models.

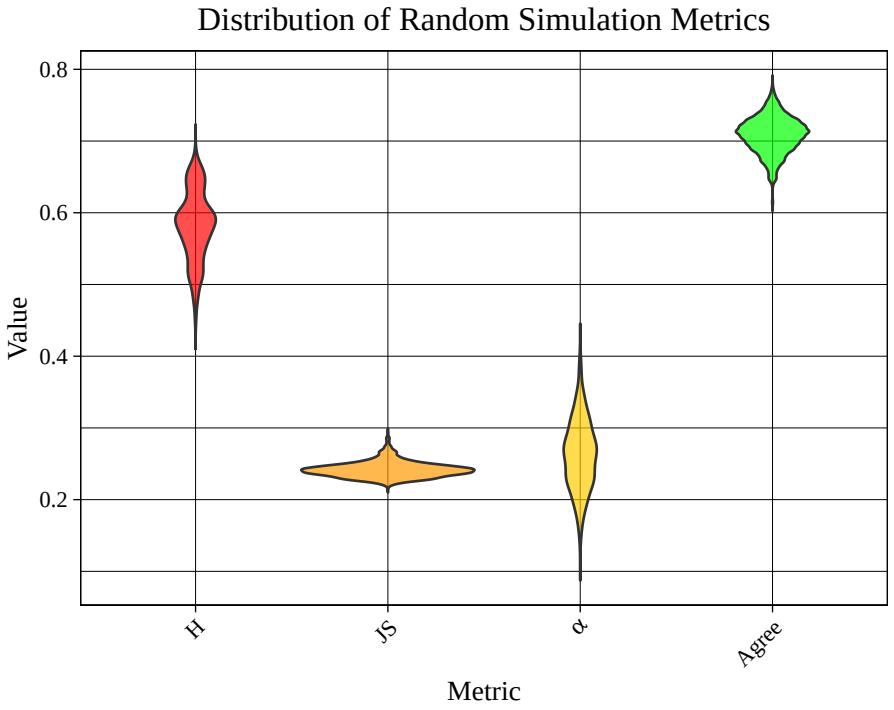

Figure 8: Metric values after 1000 random trials of RANDOMMIX (10 models).

G Text Embedding Comparison

Though many prior studies have utilized off-the-shelf text embeddings to investigate LLM explanations, Figure 11 demonstrates a clear problem with that method. Here, we show that three near-SOTA text embeddings (sampled from the top of the MTEB leaderboard) all cluster explanations by item, exhibiting a strong topic signal. This is likely due to the fact that, for a given item, all corresponding explanations will contain considerable lexical overlap — as well as other semantic similarities. As such, any two explanations *within* a given item will yield a much shorter distance to each other

Experiment	H	$JS \downarrow$	$KL_{(P\|\|Q)} \downarrow$	$KL_{(Q\|\|P)} \downarrow$	Agree ↑	α
Human Participants	0.71	-	-	-	-	0.26
MODEL MIX$_{\text{all}}$	0.63	0.22	1.28	0.78	0.72	0.26
MODEL MIX$_{\text{hi}}$	0.53	0.25	2.54	0.88	0.74	0.37
MODEL MIX$_{\text{lo}}$	0.60	0.25	2.16	0.80	0.66	0.18
`claude-haiku-3.5`	0.05	0.44	9.68	1.11	0.60	0.93
`claude-opus-4`	0.17	0.37	7.35	1.15	0.69	0.71
`command-a`	0.05	0.41	8.75	1.26	0.68	0.93
`command-r`	0.26	0.46	9.54	2.56	0.49	0.57
`deepseek-r1`	0.21	0.36	6.81	0.82	0.66	0.65
`gemini-2.0-flash`	0.04	0.45	10.31	1.48	0.57	0.95
`gemini-2.5-pro`	0.16	0.38	7.31	1.03	0.64	0.69
`gpt-4.1`	0.05	0.40	8.63	1.25	0.74	0.92
`gpt-4o-mini`	0.09	0.43	9.48	1.21	0.57	0.63
`grok-3-mini`	0.24	0.34	5.85	0.63	0.70	0.64
`grok-4`	0.16	0.39	7.69	0.93	0.64	0.64
`kimi-k2`	0.04	0.40	8.63	0.87	0.66	0.95
`llama-3.1-8b`	0.05	0.46	10.46	1.20	0.57	0.55
`llama-4-maverick`	0.08	0.49	11.32	3.51	0.44	0.89
`ministral-8b`	0.09	0.46	10.15	1.28	0.55	0.50
`mistral-medium`	0.10	0.41	8.48	0.73	0.62	0.88
`olmo2`	0.23	0.41	7.85	1.12	0.56	0.58
`phi4`	0.10	0.41	8.85	1.18	0.63	0.84
`qwen2.5`	0.07	0.42	9.21	1.05	0.60	0.90
`qwen3`	0.24	0.36	6.45	1.23	0.63	0.62

Table 5: Aggregate results across all experimental settings for every individual MODELMIX model. All metrics computed with respect to 5 individual runs, then averaged across items.

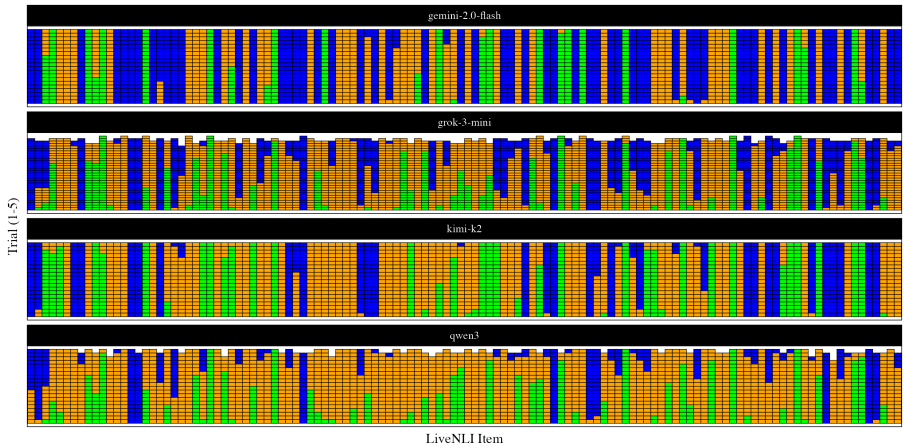

Figure 9: Model classification across trials, with varying τ (each cell represents a single classification within a τ-adjusted trial).

than to explanations pertaining to other items — even if they represent different labels or divergent reasoning patterns. As such, we opt to de-lexicalize our embeddings with respect to the premise and hypothesis text.

H Extended Explanation Variation Results

Figures 12, 13, and 14 show the results across all embedding-based metrics for the τ, top-p, and model size experiments. Though there is some slight variation across experiments, the overall trends are very similar to what is observed in Figure 4.

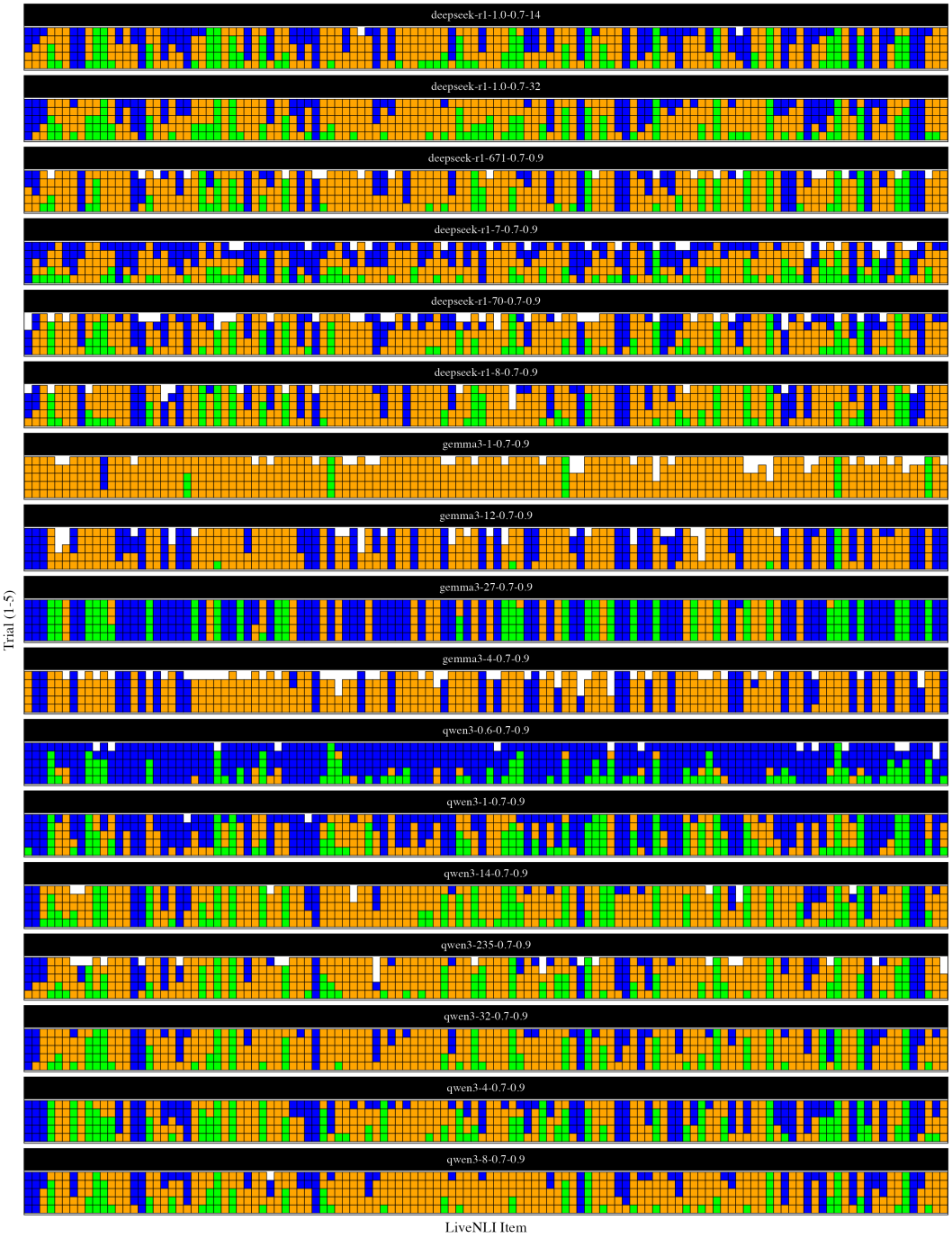

Figure 10: Model classification across trials, with varying model size (each cell represents a single classification within a trial).

Figure 11: t-SNE-reduced text embeddings across the entire LIVENLI dataset.

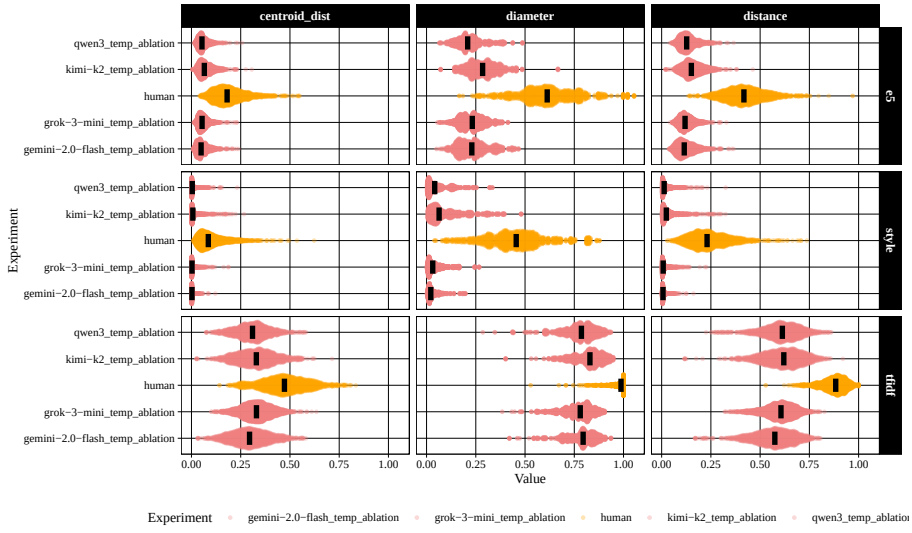

Figure 12: Explanation variation results for τ experiments.

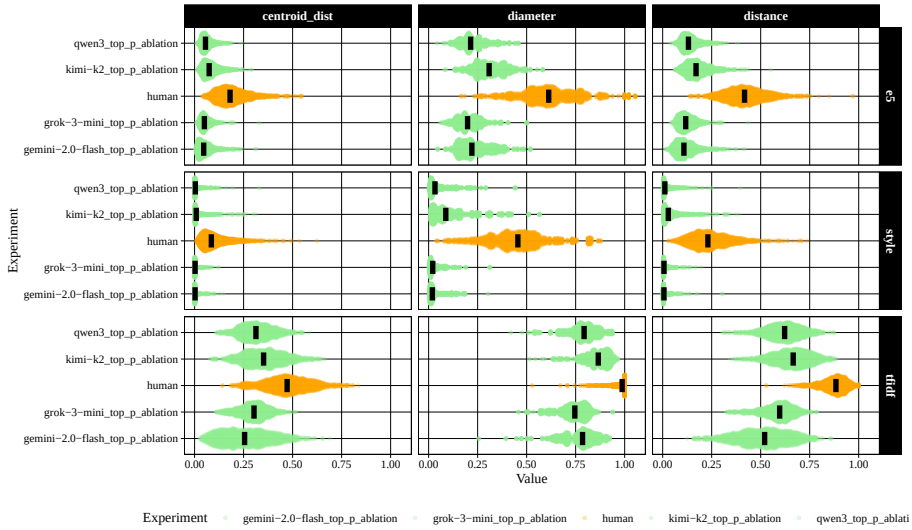

Figure 13: Explanation variation results for top-p experiments.

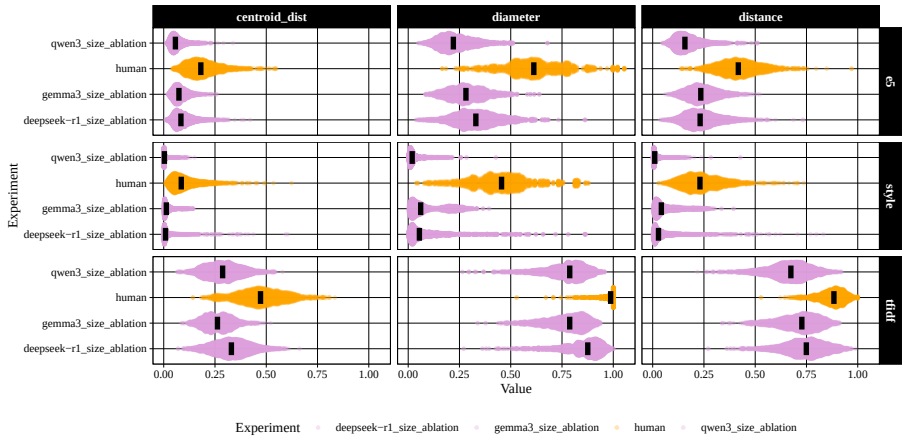

Figure 14: Explanation variation results for model size experiments.