Human-Alignment, Calibration, and Activation Patterns in Large Language Model Uncertainty

Anonymous ACL submission

Abstract

Uncertainty Quantification is a large and growing subfield of large language model behavioral analysis. Primarily to recognize and combat hallucination, the field has largely focused on measuring and improving calibration, the accuracy of uncertainty judgments to task efficacy. In this work, we investigate the relatively underexplored question of how similar large language model uncertainty is to human uncertainty. We investigate the presence and strength of human-similar uncertainty signals, deemed uncertainty alignment, in large language model overt behavior and internal activation patterns. We identify whether the models show evidence of simultaneous alignment and calibration on a variety of datasets covering both multiple choice and open ended factual recall. And we characterize the effect of instruct fine-tuning on each of these facets.

1 Introduction

Quantification and characterization of the uncertainty of transformer-based large language models (LLMs) has emerged as a major research focus. Accurate uncertainty quantification (UQ) is an important factor in the recognition and mitigation of model hallucinations (Farquhar et al., 2024) and the building and maintenance of user trust in LLM-based applications under threat of hallucination. While much research has been dedicated to the development of UQ methods and models that are well-calibrated, meaning that the uncertainty accurately indicates model efficacy (Guo et al., 2017; Shorinwa et al., 2025), very little work has investigated the related question of whether LLMs are human-like in their uncertainty behavior.

We believe the question of human-similarity in LLMs may have significant implications for human-AI collaboration. Greater levels of similarity in how uncertainty occurs and presents may foster greater rapport between LLMs and users when uncertainty judgments are provided, as human users might safely transfer their intuitive expectations of behavior under uncertainty from human-human interpersonal interaction to human-AI collaboration. Conversely, dissimilar uncertainty incidence may benefit human-AI collaboration by allowing each to cover for the other's weaknesses.

Beyond strict utility, human-similarity (or dissimilarity) in LLM uncertainty may have significant implications for the transferability of human research on behavior under uncertainty to expectations of AI behavior and vice versa. Both similarity and dissimilarity may also yield further questions about the mechanisms that underlie either, including comparisons in memory organization and information availability during learning.

In this work, we expand on the existing literature in both breadth and depth to better understand the incidence, presentation, and source of human aligned uncertainty across a diverse variety of LLMs. We start by expanding the work of Moore et al. (2026) to a larger breadth of model coverage and a set of larger and more diverse datasets. We make a first foray into measuring uncertainty alignment in a free response context. Uncertainty alignment and calibration are then tied together by measuring simultaneous alignment and calibration in identical contexts. Finally, we probe LLM activations to characterize how and where uncertainty alignment arises internally. We identify, in short, the following primary findings:

1. LLM uncertainty is generally weakly-to-moderately positively correlated with human uncertainty, mediated by task and model.
2. Both uncertainty alignment and calibration are significantly degraded by instruct-finetuning.
3. Human-aligned uncertainty is more strongly detectable in internal activations than in output logits, with qualitative differences between group-level and individual-level alignment.

2 Uncertainty Quantification

Uncertainty quantification in LLMs has been extensively studied, typically with the stated goal of detecting and reducing model hallucination. In this section, we briefly survey the existing UQ research and how calibration, the typical target of UQ research, is measured.

2.1 Prior Work

Uncertainty quantification in the general sense has grown rapidly as a subfield of LLM research. The techniques employed have grown numerous and have been well-surveyed (Shorinwa et al., 2025; Liu et al., 2025). Research in this field is typically targeted towards identifying and fostering uncertainty calibration, with relatively little work focusing on establishing human similarity. Existing work has investigated how similar the calibration of LLMs is to that of humans (Sun et al., 2025), how human estimations of LLM certainty differ from true uncertainty (Steyvers et al., 2025), and whether uncertainty-modulated human-like behaviors are also exhibited by LLMs (Xu et al., 2025; Roberts et al., 2024; Moore et al., 2025). To our knowledge, the only existing work that investigates uncertainty alignment directly is Moore et al. (2026), which found initial but weak evidence of uncertainty alignment in LLaMa and Mistral models using a small subset of the measures we here employ. Additionally, they did not investigate model activations or open response.

2.2 Expected Calibration Error

Model calibration is most commonly measured in terms of expected calibration error (ECE), defined as the expected absolute difference between a model's confidence (generally taken as the additive inverse of uncertainty) and the model's accuracy given that confidence level (Guo et al., 2017).

Because model correctness is binary per instance, this measure is typically estimated using a binned approximation. Each of the n instances are separated into a set of m bins based on confidence level, giving each bin an associated within-bin accuracy, $acc(B_i)$, and within-bin average confidence, $conf(B_i)$. The ECE is then taken to be the weighted average absolute difference between these two quantities across all bins (Equation 1).

$$ECE \approx \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B)m)| \quad (1)$$

ECE is known to be sensitive to the choice of binning procedure with both the number of bins and the across-bin probability mass distribution impacting how successfully ECE captures true calibration (Wang, 2023). In this work, all LLM calibration measurements use a technique called ECESweep (Roelofs et al., 2022). This method dynamically determines bin count such that the number of bins, m, is maximized while maintaining monotonicity of $acc(B_i)$ with increasing i for all bin counts in the range $[2, m]$. The authors find that this method effectively reduces binning bias and better detects miscalibration.

3 Methods

In this section, we describe the experimental details, covering the variety of LLMs tested (Section 3.1), the datasets used (Section 3.2), the methods used for quantifying LLM uncertainty (Section 3.3), model inference and query construction (Section 3.4), and activation probing techniques employed (Section 3.5).

3.1 Model Choice

We use 30 open-weight text-only LLMs representing a variety of model families (covering LLaMa, Mistral, Gemma, and Falcon) and models sizes (ranging from 1B to 13B parameters per model). Specifically, we tested LLaMa 2 (7B and 13B) (Touvron et al., 2023), LLaMa 3 (1B, 3B, and 8B) (Grattafiori et al., 2024), Mistral (0.1 and 0.3, both 7B) (Jiang et al., 2023), Gemma (2B and 7B) (Team et al., 2024a), Gemma (2B and 9B) (Team et al., 2024b), and Falcon 3 (1B, 3B, 7B, and 10B) (Team, 2024). In all cases, we tested both the instruct-finetuned and base versions of every model.

3.2 Datasets

All non-probing experiments were performed identically over four datasets. All datasets are question answering tasks with substantial variation in topic coverage and question format.

Our first dataset is the commonly employed multiple choice question answering (MCQA) MMLU dataset (Hendrycks et al., 2020). This dataset consists of more than $14K$ multiple choice questions drawn from a wide variety of factual subjects. As this dataset provides no information regarding per-question human uncertainty response, it is included primarily as an established baseline when measuring ECE for the remaining datasets.

Each of the remaining datasets provide human response data from which we derive estimates of per-question human uncertainty. The first is ProtoQA (Boratko et al., 2020), which is a Family Feud inspired dataset consisting of 8724 crowd-sourced purely preference or opinion-based questions. Each question has a variable-size set of answers provided by human survey respondents. On average, each question is associated with 5.033 answer choices and 89.407 independent human respondents. The second dataset, CamChoice (Mullooly et al., 2023) comprises 504 reading comprehension questions where each question is associated with a passage on which the question is based. All questions in this dataset have exactly 4 available answer choices and each passage is associated with an average of 6.46 questions. During inference, passages are prepended to questions such that each question and its associated passage are presented to the model in isolation. For both the CamChoice and ProtoQA datasets, we have access to the percentage of human respondents that selected each of the available answer choices. For these datasets, we took the human uncertainty to be the normalized entropy over these response percentages, which we consider a measure of *group uncertainty*.

The final dataset is named herein the Coane dataset, after the first author of Coane and Umanath (2021). This dataset consists of 421 general knowledge questions that were presented to older adult subjects. Like ProtoQA and CamChoice, the Coane dataset provides the percent of subjects who selected each of 4 answer choices per question, allowing the computation of entropy over these responses as well. In addition, Coane also provides the average response time. They separate response times based on correctness, but provide sufficient information to reconstruct the average response time across all subjects, which we consider a measure of *average individual uncertainty* following the established interpretation in cognitive science (Hick, 1952). In addition to the MCQA task, Coane also provides human response data on an open ended question answering (OEQA) variant over the same question set. This allows us to investigate uncertainty alignment in both MCQA and OEQA contexts without substantial differences in context. The human data provided allows us to calculate the average response time over all respondents, which we interpret as the average individual uncertainty in a OEQA setting.

3.3 Uncertainty Measures

Our experiments are limited to a wide selection of inference-time LLM UQ measures that can be calculated from the instantaneous model logits and activations at each inference timestep. Such per-iteration uncertainty signals, in addition to being computationally cheaper than multi-inference UQ methods, are more conducive to signal utilization in downstrean applications and analysis. For example, future analysis regarding uncertainty incidence throughout generation would be difficult or impossible with only access to sequence-level uncertainty values. To facilitate broad coverage under possibility of measure choice-senstivity and coverage of both MCQA and OEQA tasks, we experiment with 4 broad but interconnected categories of UQ measures: MCQA measures, Free Response (FR), Bracketed FR (BFR), and 1st token FR (1TFR).

3.3.1 MCQA measures

The basic set of uncertainty measures are defined for MCQA and are largely adapted from those used in Moore et al. (2026). These measures fall into 5 subcategories: full distribution measures, cloze measures, top-k entropies, top-p entropies, and top-p sizes. There are two full distribution measures: top-1-prob and total-ent. The former is simply the additive inverse of the highest probability token across the entire vocabulary, that is $1 - \max_{v \in V}(P(v|C_{MC})$ where C_{MC} is a multiple-choice formatted context and V is the full model vocabulary. The additive inverse is taken because higher probability mass on the top token suggests higher certainty, therefore lower uncertainty. The other measure, total-ent, is the normalized entropy over the full vocabulary set (Equation 2). As probability mass becomes more diffuse across the vocabulary, suggesting that either more tokens are taken as valid alternatives or the most preferred token(s) become less clearly preferred, the entropy of the distribution will rise. This intuitively suggests entropy over outcomes (vocabulary tokens) as a valid measure of model uncertainty. This is established in the literature (Huang et al., 2025b), but is relatively uncommonly used in favor of more complex methods.

$$-\frac{1}{\ln|V|}\sum_{v \in V} P(v|C_{MC}) \ln(P(v|C_{MC}) \quad (2)$$

All remaining measures are calculated based on various focused subsets of the total probability dis-

tribution. These are chosen based on expectation that the probability mass and the meaningful uncertainty signal will typically be concentrated in a small portion of the overall vocabulary. Focusing exclusively on these vocabulary subsets may allow identification of signals that would otherwise be lost in the full distribution measures or better discernment between uncertainty and low entropy caused by task-irrelevant syntactic or semantic restriction of the probable token set. Most of these remaining measures are defined as the entropy over three focus subsets: cloze, top-k, and top-p. In all cases, the entropy is calculated exactly as in Equation 2 with two major changes.

First, the vocabulary, V, is replaced by a strict subset V_k, V_p, or V_c. V_k represents the k highest probability tokens from $P(V|C_{MC})$. V_p is the top-p tokens, meaning the minimum set of tokens such that $\sum_{v \in V_p} P(v) \geq p$. V_c is a pre-defined set of tokens that represent the set of valid cloze test completions for the MCQA context.

To handle semantically equivalent variations, probabilities are summed across variations of valid cloze tokens. For example, the probability of the cloze target 'A' is $P(A \in V_c) = P(\text{'}A\text{'}) + P(\text{'}A)\text{'}) + P(\text{'}A.\text{'}) + P(\text{'}a\text{'}) + \ldots)$. The probability of the tokens within the subsets are renormalized to sum to 1 by dividing by the total subset probability, that is $P(v^i \in V_i) = P(v \in V) / \big(\sum_{v \in V_i} P(v)\big)$.

For top-k and top-p entropies, uncertainty is measured across a range of reasonable subset sizes, specifically $k \in \{5, 10, 25, 50, 100\}$ and $p \in \{0.95, 0.90, 0.75, 0.50\}$.

The remaining measures are not entropy based. The first is the top-1 probability over the cloze token set, V_c (after the previously described semantic grouping). Ideally, but not necessarily, the most likely token given an MCQA-formatted context will be one of the valid tokens within the cloze token target set. This yields an expectation, but not a guarantee, that top-1-prob is less than but nearly equal or proportional to choice-prob. The remaining measures are derived from the top-p subset, V_p. We take the number of tokens that fall into this set $|V_p|$ to be a measure of uncertainty intuitively representing the number of valid completion tokens given C_{MC}. This measure is alternatively known in the literature as the nucleus size and is connected to the credible interval, which is the Bayesian analog of the commonly employed confidence interval.

3.3.2 FR, BFR, and 1TFR Measures

The various free response uncertainty measures are largely derived directly from the MC measures described in the previous section. The only significant exception to this is sequence perplexity (PPL). Perplexity, defined as used herein in Equation 3 where N is the number of generated tokens and W is the generated token sequence, is a common metric of sequence probability and surprisal. While PPL is typically defined over the entire sequence, including both starting context and model response, we take the perplexity over only the generated sequence to avoid confounding effects from relative likelihood across varying contexts. Being analogous to sequence-level probability, this measure was also converted to uncertainty by taking its additive inverse, $1 - PPL(W|C_{FR})$.

$$PPL(W|C_{FR}) = exp(-\frac{1}{N}\sum_{i=1}^{N} \ln P(w_i|C_{FR}, w_1, \ldots, w_{i-1})) \tag{3}$$

The remaining FR measures are simple averages over all generation steps of the non-cloze measures as defined in the previous section. This includes the generated sequence average total entropy, top-k entropy, top-p entropy, and top-p size. The entropy measures deserve some additional justification for this choice. Entropy over the probability distribution is defined as the expected log surprisal of each outcome. Because the resulting value is therefore in logspace, the arithmetic mean of entropy is equivalent to the geometric mean of the expected surprisal. As surprisal is itself in probability-space, this results in the surprisal that, if held constant across all timesteps, would yield the observed sequence surprisal, making it an effective length-normalized measure of total sequence surprisal.

The remaining BFR and 1TFR uncertainty measures are simple to define in terms of MC and FR measures. As described in further detail in Section 3.4, our OEQA prompts are designed to encourage the model to output the selected answer immediately and mark the end of the chosen answer with a closing '}' brace. The intention is to identify the most relevant answer segment so that further model explanation or otherwise over-generation can be excluded from UQ measurement. This is an especially important consideration for models

that have not been instruct finetuned, as they often fail to generate a stop token, opting instead to generate additional questions until the maximum generation length is reached. BFR is exactly the base FR measures applied exclusively to the sequence segment that the model generated before the closing brace. We also hypothesized that most of the alignment signal may be further concentrated in the first generated token. We test this by measuring 1TFR measures, which is exactly the non-cloze MC measures applied to the first token generated by the model during OEQA. All three FR measure sets are calculated on the same generated output without regeneration.

3.4 Model Inference

MCQA and OEQA prompts were designed similarly. The models were presented with two ICL examples, structured appropriately for MCQA or OEQA as necessary. Prior to the ICL examples, both templates also included a short system prompt, either in the designated system prompt location for supporting models or at the beginning of the context window, with instructional content to encourage models to answer the question immediately, without explanation, and ending with a closing brace to facilitate the BFR measures. After presenting the question, and answer choices for MCQA, the model's answer was artificially started with appropriate formatting for chat-finetuned models such that the first generated token is expected to contribute directly towards the intended answer rather than superfluous formatting. While this nudging might limit the generalization of the results herein to more complex tasks, it has the added benefit of better isolating to the uncertainty behavior on specifically the tokens that contribute to the answer without noise from framing tokens.

Model inference was performed using 5 A100 GPU hours in a container with identical methods across all MCQA datasets with minimal adaptation to accommodate OEQA for the Coane dataset. Where possible, each model's starting context was formatted in line with trained expectations as defined in the associated tokenizer chat template as provided by the huggingface apply_chat_template method. For models that support system prompts, the instructional content portion of the chat template was instead included in the system prompt rather than the main prompt body. Model-specific template preparations were also applied the ICL examples to maximize consistency.

All contexts were dynamically batched after formatting such that all contexts within a given batch were identical in terms of tokenized length. This ensured that no contexts included padding tokens which might have affected absolute token probabilities. For OEQA tasks, padding tokens may have been generated with batches yielded differing generation lengths, but padding tokens were stripped before UQ measure calculation.

Model correctness, necessary for downstream ECE calculation, required substantially different methods between the MCQA and OEQA tasks. MCQA correctness was readily obtainable with the model's chosen answer is taken to be the answer choice to which the highest total probability is assigned among the available answer choices. This chosen answer label is then compared against the expected answer label to determine correctness. To account for synonymous variations in label (e.g. "A" vs. "A.", "A)", "a", etc.), we sum across a collection of reasonable synonymous variants such that the probability of answer A is taken the be $P(\text{`}A\text{'}) + P(\text{`}A.\text{'}) + P(\text{`}A)\text{'}) + P(\text{`}a.\text{'}) + \ldots$ and similarly for all MCQA answer labels.

OEQA correctness was instead judged manually for correctness individually by two human annotators by comparing model generations against the expected answer, with disagreements between annotators being resolved by a third annotator. Reasonable allowances were made for spelling, grammar, synonymy, and correct alternatives for underspecific questions. During this process, two questions in the Coane dataset were identified as unsuitable for inclusion. Details and justification on these two are briefly presented in Appendix A. These two questions were excluded from all downstream analysis, with the exception of the model probing described in Section 3.5, due to the late discovery. The small number of affected questions is unlikely to have substantially affected the final results, however.

3.5 Model Probing

To investigate where and how uncertainty alignment presents in the LLM internals, we use a probing method inspired by Bürger et al. (2024), in which we train linear regression models that attempt to predict human uncertainty from LLM activation at each layer. For a given LLM, we presented the Coane MCQA formatted contexts to the LLM and extracted the last-token activations for each context. We used the resulting activations, to train

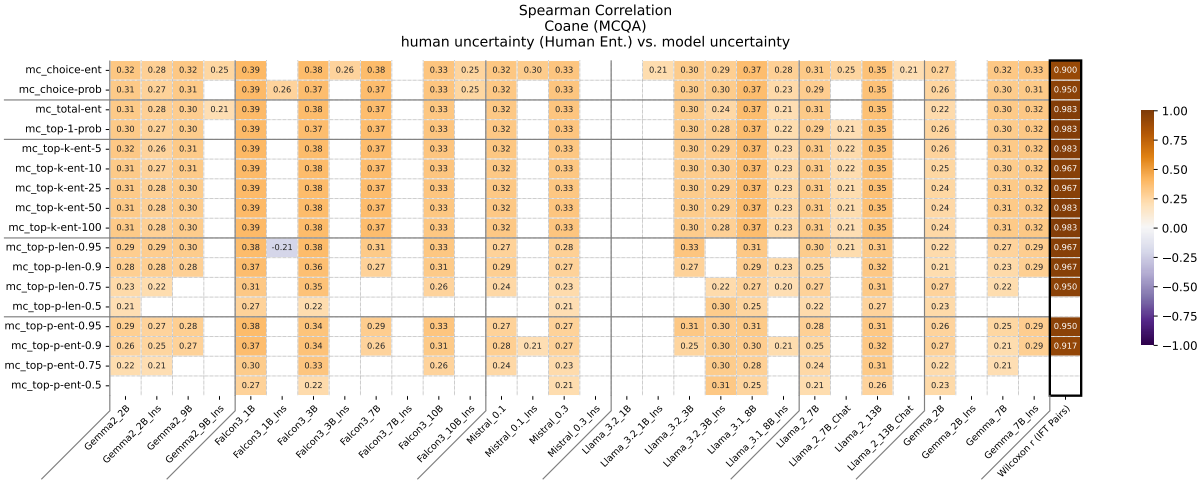

Figure 1: Spearman correlation between uncertainty measures and human uncertainty (measured via human response distribution entropy) on the MCQA version of the Coane dataset. Darker cells indicate higher correlation, while blank cells indicate that correlation failed to meet the corrected p-value threshold. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

a simple linear regression model for each layer of the LLM with human uncertainty as the regression target. A separate model was trained for each of three measures of human MCQA uncertainty: response distribution entropy, average response time when correct, and average overall response time. We employed 10-fold cross validation when training each layer's linear regressors and evaluated the regressors using Pearson r correlation.

4 Results

In this section we present the results of our various analyses.

4.1 Agreement and Correctness

As an initial surface-level analysis, we calculated for each model, the model's correctness on factual datasets (Coane, CamChoice, MMLU) and the model's agreement with human answer preference for datasets with human response distribution data (Coane, CamChoice, ProtoQA). Model correctness on factual tasks scaled positively with parameter count in base and instruct models (r = 0.45–0.71 across factual datasets), with instruct-finetuned variants showing no systematic advantage over their base counterparts except in open-ended factual recall, where finetuning was associated with a significant decrease in accuracy (Wilcoxon p = 0.017, r = 0.67). Agreement with the most common human response followed a similar but attenuated pattern on knowledge-based datasets, and was largely independent of model size on the purely preference-based ProtoQA dataset

Given that no such size-dependence is discernable for the purely preference-base ProtoQA, however, this is likely only an artifact of size-dependent correctness and human accuracy. Generally, agreement per model is ranked such that correctness on MCQA tasks is higher than than both OEQA correctness and MCQA human-similarity. OEQA correctness tends higher than MCQA human-similarity for all nearly all models larger than 3B.

4.2 Uncertainty Alignment

The rightmost column of Figure 1 displays the significant effect sizes per measure for Wilcoxon signed rank tests between the alignment of each base model versus the alignment of its instruct-finetuned counterpart. Across nearly every measure, this test finds a strong relationship such that instruct finetuned models tend to be less aligned that their base counterparts.

The remaining results for other all other dataset, task, and human uncertainty configurations can be found in Figures 4-9 in Appendix C. In brief, the noteworthy trends are that the alignment signal is present but far more sporadically and model-dependent for the CamChoice dataset and for MCQA response time correlation on Coane. ProtoQA shows consistently significant, but very weak alignment. For OEQA, the models show nearly no significant evidence of alignment for any of the

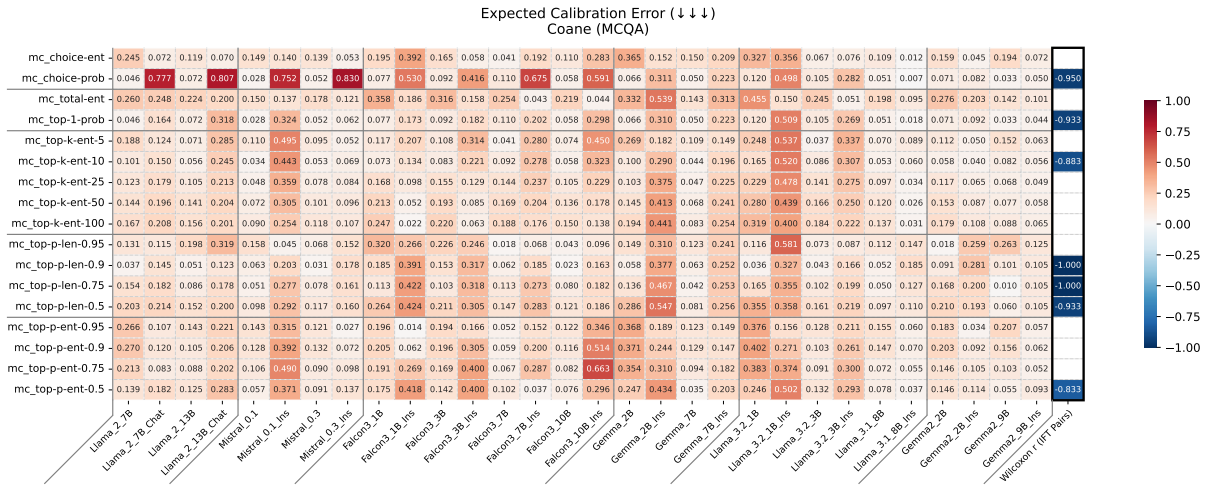

Figure 2: ECESweep results for the MCQA version of the Coane dataset. Darker cells indicate higher calibration error and thus lower calibration. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models, with blank cells indicating insignificance. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

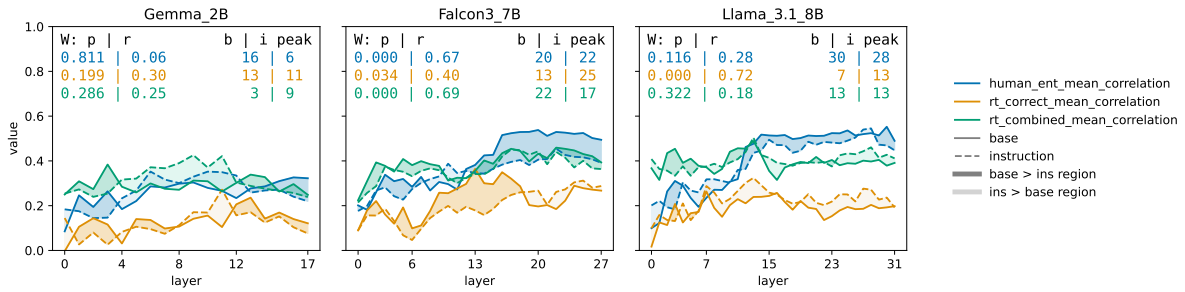

Figure 3: Per-layer model activation probing correlations. Colors indicate the human uncertainty type targeted during linear regression. Shaded regions indicate the gap between the base mode (solid line) and the instruct finetuned variant (dashed line). On-figure annotations in the top-left indicate the p-value and effect size for a Wilcoxon rank sum test between the base correlation and ins correlation across layers. Annotations in the top-right display the layer at which maximum human alignment was measured.

FR uncertainty measure classes. Only Falcon 3 10B shows any significant correlation with human response time, but this may be an artifact of extraneous generation behavior as the same model shows no alignment for BFR or 1TFR.

Unlike factual accuracy, uncertainty alignment shows no consistent relationship with model size in either base or instruct models.

4.3 Uncertainty Calibration

Following up on its standout alignment in the prior section, we have evaluated the ECE scores for each MCQA measure-model pair on the Coane dataset in Figure 2, with the remaining datasets and Coane OEQA measures presented in Figures 10-14 in Appendix C. With a few noteworthy exceptions, the results show relatively strong calibration on the Coane dataset, with every model showing strong calibration (ECE $\lesssim 0.1$) for at least one measure. Like with alignment, there is noticeable model-dependency, with the Mistral 0.1 Ins, Gemma 2B Ins, and LLaMa 3.2 1B Ins showing substantially worse calibration across nearly every measure than most other models. For measures, mc_choice-prob stands out as particularly volatile, being highly calibrated for most models, but actively anti-calibrated for the instruct-finetuned variants of all LLaMa 2 and Mistral models. The Wilcoxon column shows evidence of significant degradation in calibration after instruct finetuning across many measures. This result, while initially surprising, is consistent with existing findings on the unintended effects of instruct finetuning (Huang et al., 2025a). Taken together with the results of section 4.2, this suggests that alignment and calibration are both negatively affected by instruct finetuning.

Unlike alignment, calibration appears less affected by choice of model and dataset, with lower but still substantial calibration across all pairings for both MMLU and CamChoice. This may be explainable by the relative difficulty of these two datasets when compared with Coane MCQA. Unlike the alignment results, the OEQA results on Coane show strong calibration across all three uncertainty measure groups. Average top-k entropy in particular shows strong calibration for full generation measures with relatively little loss in BFR. In a noteworthy deviation from the MCQA case, we observe a significant positive Wilcoxon effect size for some measures, indicating some measures may improve calibration with instruct finetuning outside of limited-option contexts.

As a final point of comparison, we calculated the human calibration on the Coane dataset. Because we have access to per-question accuracies, we are able to sidestep binning questions by taking average instance-level calibration $\frac{1}{|Q|}\sum_{q \in Q} |Acc(q) - Conf(q)|$ where Q is the set of all questions. The results were average calibration errors of 0.226 for MCQA with human entropy, 0.210 for MCQA with response time, and 0.445 for OEQA with response time. This indicates that the LLMs tend to be more well-calibrated than human subjects across most UQ measures.

4.4 Probing

The results of probing the LLM internals are shown in Figure 3 for a representative subset of models, with the remaining models reported in Figure 15 in Appendix C. In addition to plotting the correlation across layers, we perform a Wilcoxon signed-rank test on the per layer base model correlation vs instruct model correlations. We also identify the layers at which maximum correlation is attained with each of the human uncertainty measures.

Consistently across nearly every model, the most correlated human measure with model activations, both in absolute terms over all layers and in terms of final layer activations, is human entropy. This dominance tends to arise near the middle layers. Notably, this occurs because of the relatively flat trajectory of the response time across layers, while human entropy tends to become more correlated with depth. This suggests a meaningful difference in how the models internalize human individual-level (RT) vs group-level (entropy) uncertainty. Group uncertainty alignment consistently occurs stronger, later, and with increasing strength throughout inference. We also see, when comparing against the alignment heatmaps in Section 4.2, that maximum correlation for every model tends to be substantially higher than that observed only in the output logits. For an extreme example, Gemma 2 9B shows a maximum of $r \approx 0.6$ correlation with human entropy when using the internal probe, but a maximum of $r = 0.32$ correlation for logit-based measures. This shows promise for further behavioral identification and isolation in future LLM behavioral studies.

5 Conclusion

In this work, we have extensively studied the uncertainty behavior of an array of pre-trained LLMs. We have verified and extended previous results by finding evidence of a consistent, but moderate in strength and sensitive to measurement methodology, positive uncertainty alignment signal. We have demonstrated that models are often simultaneously aligned and calibrated in their uncertainty and that both of these tend to be disrupted by instruction finetuning. We also showed through activation probing that LLM overt behavior does not fully capture the model alignment and found initial evidence that human similarity may be better characterized in terms of similarity to groups of humans rather than representative individuals.

6 Future Work

As the field of uncertainty alignment in LLMs remains largely unexplored, there are many viable paths for further exploration in future work. This work has remained limited to the narrow tasks of short-form question answering. Future work should expand into more complex tasks, such as problem-solving, planning, or long-form response tasks. To bring our work more in line with existing calibration research, future work should also investigate multi-inference and other more complex methods of uncertainty quantification. Finally, with the recognition of uncertainty alignment signals within model activation patterns, isolation of the related latent vectors may allow for direct manipulation, enhancement, or even reversal of alignment behaviors without significant retraining, broadening to possible applications of the research presented herein.

Limitations

The largest limitation of the study herein is the relatively limited task set, covering only simple factual recall tasks. It is difficult to guarantee generalization of any results reported herein to more complex tasks, especially in light of the nearly nonexistent overt alignment results on the OEQA task. Our most well-performing dataset, Coane, is also targeted at an age-specific cohort of older adults. This may harm generalization to more diverse human comparison sets, though it is also possible that the knowledge biases within this group may have artificially deflated the final results.

References

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. Protoqa: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. *Advances in Neural Information Processing Systems*, 37:138393–138431.

Jennifer H Coane and Sharda Umanath. 2021. A database of general knowledge question performance in older adults. *Behavior Research Methods*, 53(1):415–429.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

William E Hick. 1952. On the rate of gain of information. *Quarterly Journal of experimental psychology*, 4(1):11–26.

Jerry Huang, Peng Lu, and Qiuhao Zeng. 2025a. Calibrated language models and how to find them with label smoothing. *arXiv preprint arXiv:2508.00264*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025b. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

Kyle Moore, Jesse Roberts, Thao Thi Minh Pham, and Douglas Fisher. 2025. Chain of thought still thinks fast: Apricot helps with thinking slow. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Kyle Moore, Jesse Roberts, Daryl Watson, and Pamela Wisniewski. 2026. Investigating human-aligned large language model uncertainty. *The International FLAIRS Conference Proceedings*, 39(1).

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark JF Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, and 1 others. 2023. The cambridge multiple-choice questions reading dataset.

Jesse Roberts, Kyle Moore, Douglas Fisher, Oseremhen Ewaleifoh, and Thao Pham. 2024. Large language model recall uncertainty is modulated by the fan effect. In *Proceedings of the 28th conference on computational natural language learning*, pages 303–313.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. 2022. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231.

Fengfei Sun, Ningke Li, Kailong Wang, and Lorenz Goette. 2025. Large language models are overconfident and amplify human bias. *arXiv preprint arXiv:2505.02151*.

Falcon-LLM Team. 2024. The falcon 3 family of open models.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Cheng Wang. 2023. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*.

Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2025. Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25655–25672.

A Dataset Alterations

In this section, we briefly detail minor alterations and/or exclusions made to the datasets. The majority of changes were small encoding changes, specifically the replacement of some Unicode punctuation characters with their ASCII equivalents. This was done to guard against interference in model behavior from the inclusion of equivalent but much rarer tokens.

Most notably, two questions were removed from the Coane dataset. The first of these concerned the current holder of a public political office. This was removed for two-fold reasons. First, the correct answer to this question has changed since the original human trials, meaning that the correct answer as marked is outdated. While this could have been rectified, we removed the question entirely because the wide range of knowledge cutoffs across models makes it very likely that the "correct" accessible answer varies widely. The second question concerned a detail to a popular fantasy book, "The Little Prince". On further research, the annotators concluded that not only is the answer provided in the original dataset not correct based on publicly accessible information on the story, but the question asked has no non-trivial correct alternative answer. As such, it was also removed from all further analysis.

B Model Prompt Examples

This section will contain example prompts, including question and answer choices where appropriate, as they were presented to the models. The specific formats presented are those used for the Falcon 3 1B base model. In all cases, underscores (_) indicate the model's first token generation position and are not included in the context provided to the mode.

B.1 ProtoQA

```
Question: What is the chemical symbol for
gold?
A. Ag
B. Mo
C. Au
D. Gd
Answer: C

Question: In what year did the First World
War end?
A. 1939
```

B. 1918
C. 1914
D. 1945
Answer: B

Question: At The Beach, Name Something That Might Protect You From Sun.
A. umbrella
B. sunscreen
C. sun hat
D. sunglasses
E. cover up
F. shade
Answer: _

B.2 CamChoice

Question: What is the chemical symbol for gold?
A. Ag
B. Mo
C. Au
D. Gd
Answer: C

Question: In what year did the First World War end?
A. 1939
B. 1918
C. 1914
D. 1945
Answer: B

Question: Some time ago a website highlighted the risks of public check-ins - online announcements of your whereabouts.
...
<story truncated for example purposes only>
...
What Armstrong has figured out is something we would all do well to remember: the web may allow us to say anything, but that doesn't mean we should. Why does the writer describe a website about public check-ins in the first paragraph?
A. to reinforce the concerns already felt by some people
B. to remind readers to beware of false promises
C. to explain that such sites often have a hidden agenda
D. to show that the risks of internet use are sometimes over estimated
Answer: _

B.3 MMLU

Question: What is the chemical symbol for gold?
A. Ag
B. Mo
C. Au
D. Gd
Answer: C

Question: In what year did the First World War end?
A. 1939
B. 1918
C. 1914
D. 1945
Answer: B

Question: Find the degree for the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q.
A. 0
B. 4
C. 2
D. 6
Answer: _

B.4 Coane

B.4.1 MCQA

Question: What is the chemical symbol for gold?
A. Ag
B. Mo
C. Au
D. Gd
Answer: C

Question: In what year did the First World War end?
A. 1939
B. 1918
C. 1914
D. 1945
Answer: B

Question: Which band was Paul McCartney a member of?
A. The Monkees
B. The Rolling Stones
C. The Yardbirds

```
D. The Beatles
Answer: _
```

B.4.2 OEQA

```
What is the chemical symbol for
gold?
Answer in as few words as possible,
wrapping the key part in {} brackets:
The chemical symbol for gold is {Au}.

In what year did the First World
War end?
Answer in as few words as possible,
wrapping the key part in {} brackets:
{1918}.

Which band was Paul McCartney a
member of?
Answer in as few words as possible,
wrapping the key part in {}
brackets: {_
```

C Additional Figures

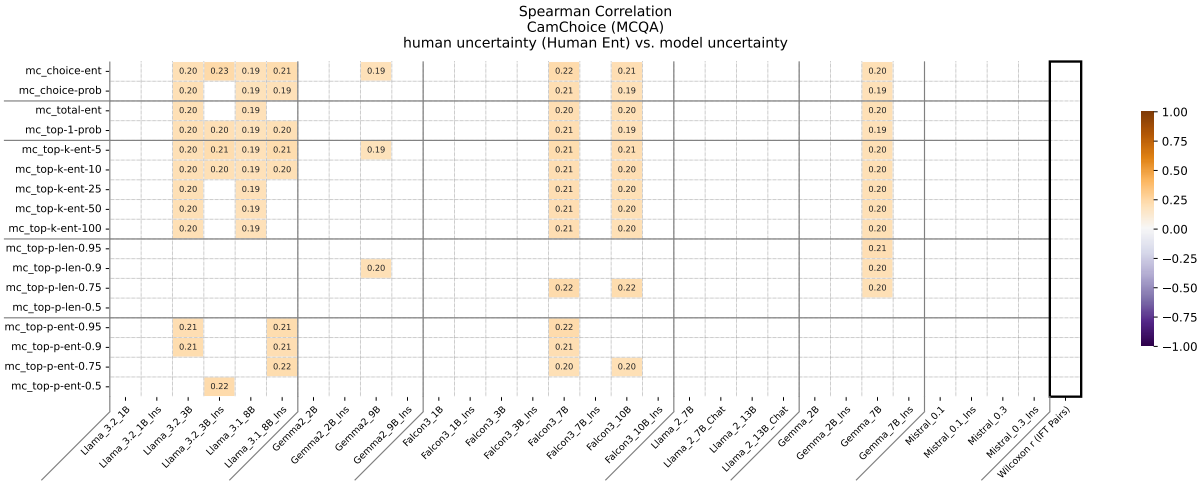

Figure 4: Spearman correlation between uncertainty measures and human uncertainty (measured via human response distribution entropy) on the CamChoice dataset. Darker cells indicate higher correlation, while blank cells indicate that correlation failed to meet the corrected p-value threshold. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

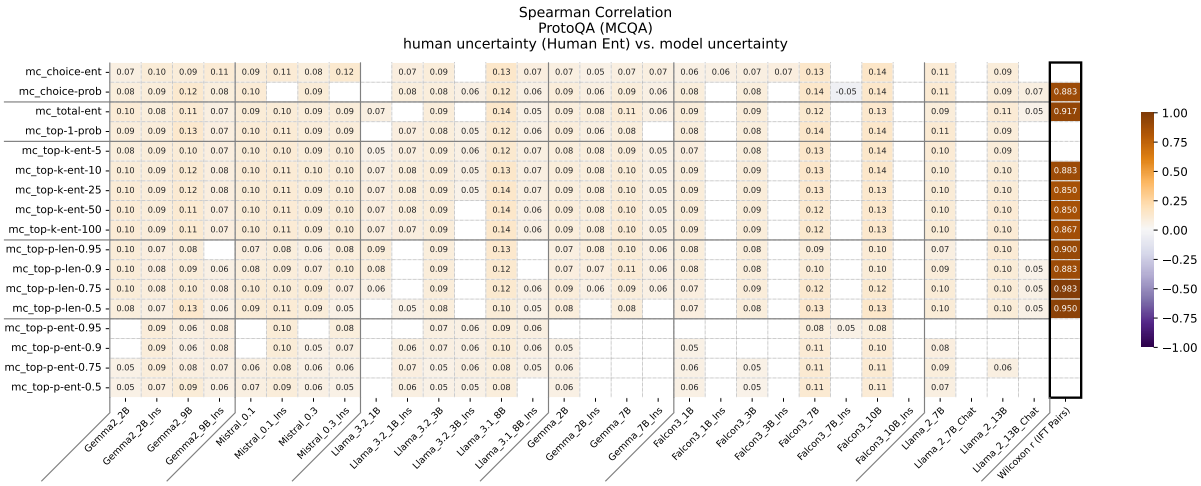

Figure 5: Spearman correlation between uncertainty measures and human uncertainty (measured via human response distribution entropy) on the ProtoQA dataset. Darker cells indicate higher correlation, while blank cells indicate that correlation failed to meet the corrected p-value threshold. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

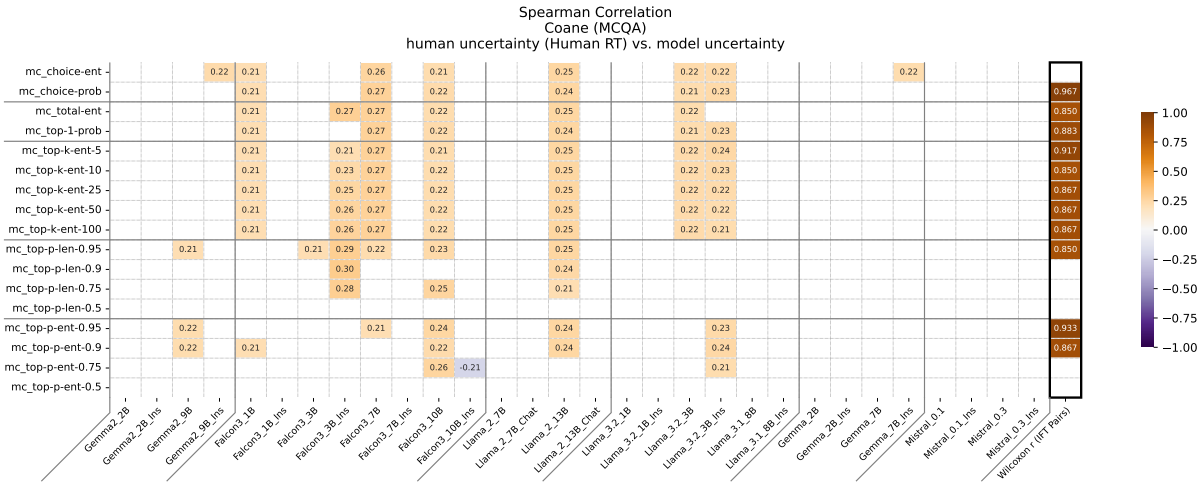

Figure 6: Spearman correlation between uncertainty measures and human uncertainty (measured via human response distribution entropy) on the MCQA version of the Coane dataset compared against human response time. Darker cells indicate higher correlation, while blank cells indicate that correlation failed to meet the corrected p-value threshold. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

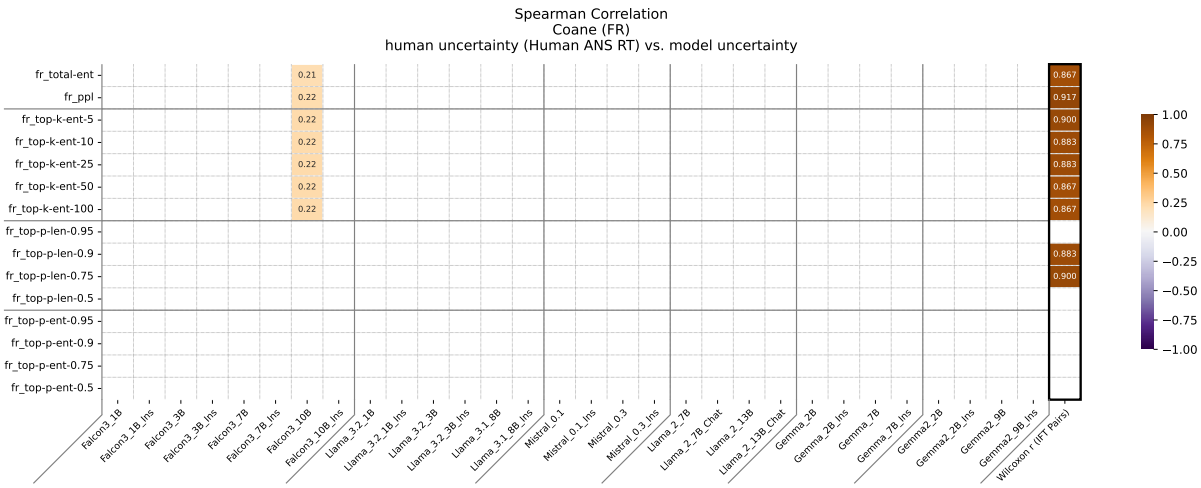

Figure 7: Spearman correlation between uncertainty measures and human uncertainty (measured via human response distribution entropy) on the FR version of the Coane dataset. Darker cells indicate higher correlation, while blank cells indicate that correlation failed to meet the corrected p-value threshold. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

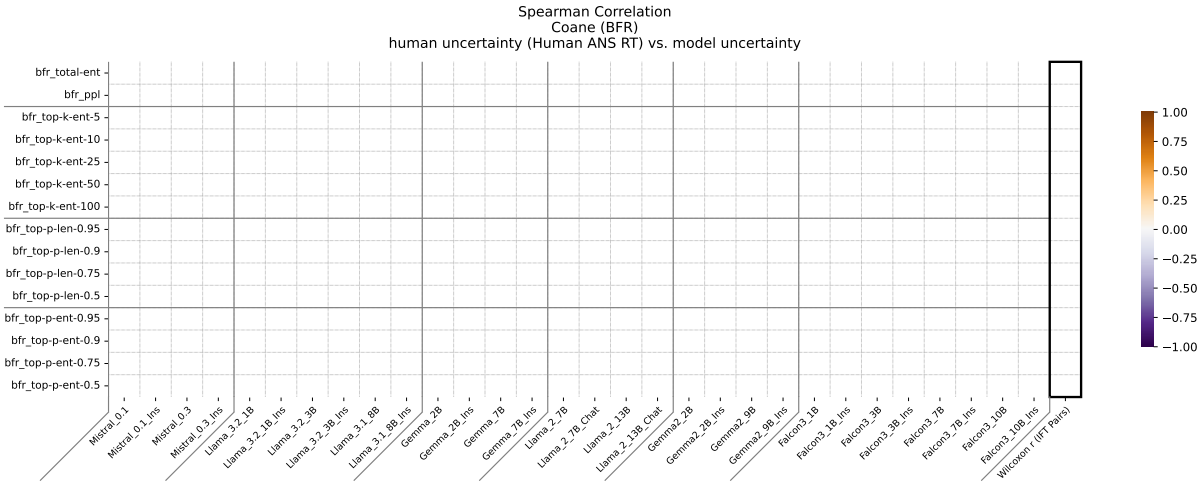

Figure 8: Spearman correlation between uncertainty measures and human uncertainty (measured via human response distribution entropy) on the BFR version of the Coane dataset. Darker cells indicate higher correlation, while blank cells indicate that correlation failed to meet the corrected p-value threshold. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

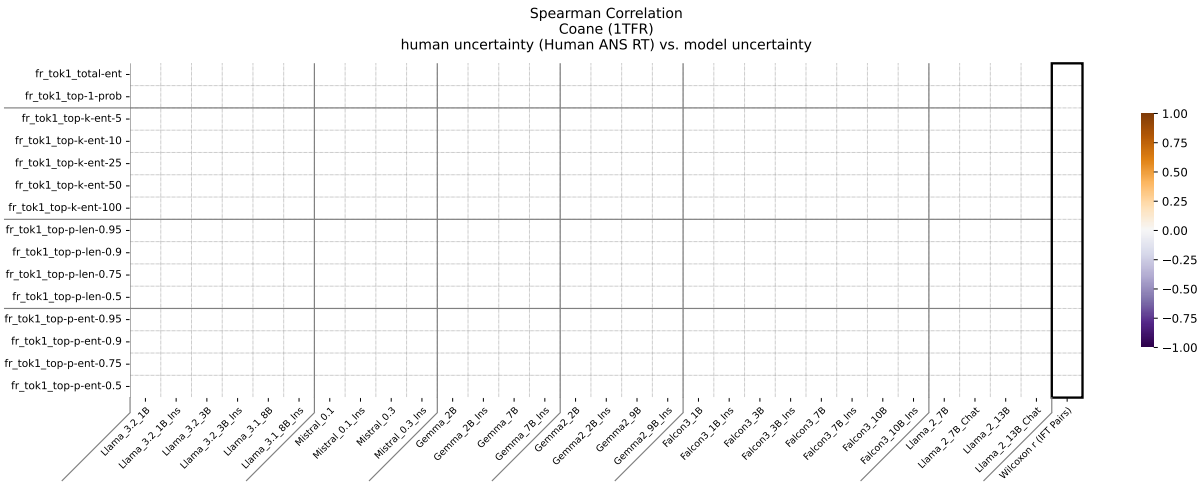

Figure 9: Spearman correlation between uncertainty measures and human uncertainty (measured via human response distribution entropy) on the 1TFR version of the Coane dataset. Darker cells indicate higher correlation, while blank cells indicate that correlation failed to meet the corrected p-value threshold. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

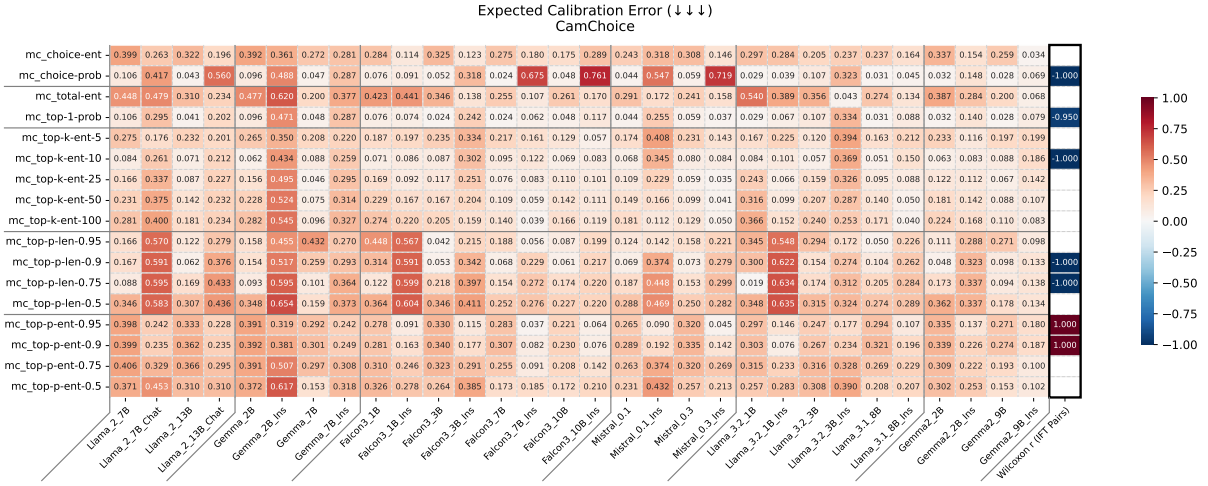

Figure 10: ECESweep results for the CamChoice dataset. Darker cells indicate higher calibration error and thus lower calibration. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models, with blank cells indicating insignificance. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

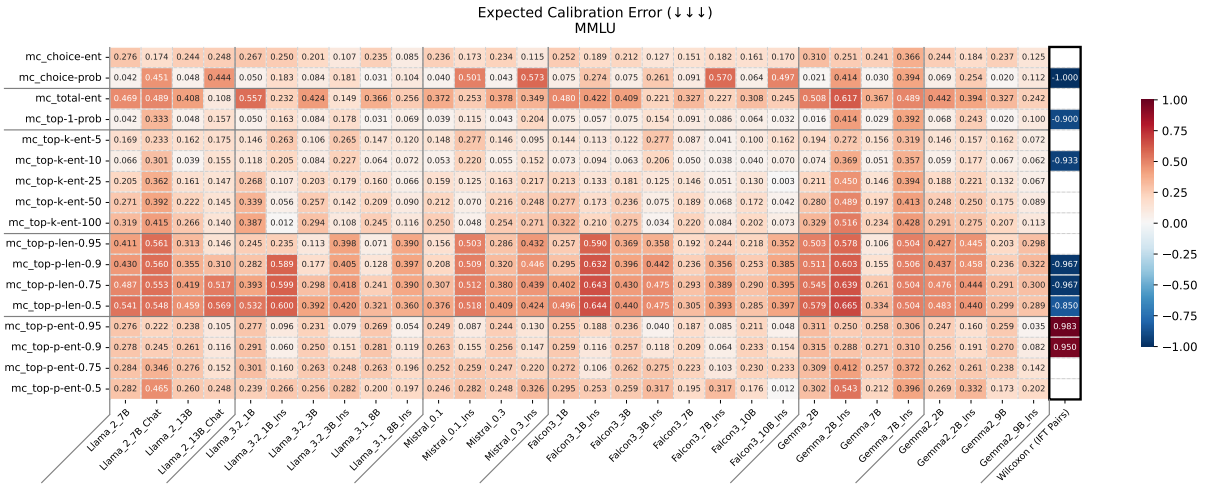

Figure 11: ECESweep results for the MMLU dataset. Darker cells indicate higher calibration error and thus lower calibration. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models, with blank cells indicating insignificance. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

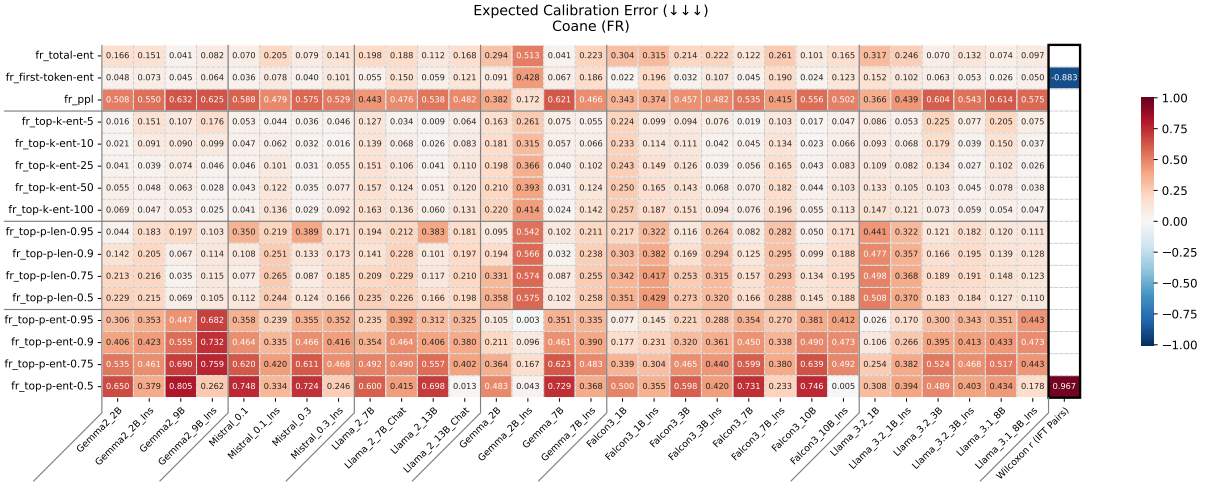

Figure 12: ECESweep results for the FR version of the Coane dataset. Darker cells indicate higher calibration error and thus lower calibration. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models, with blank cells indicating insignificance. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

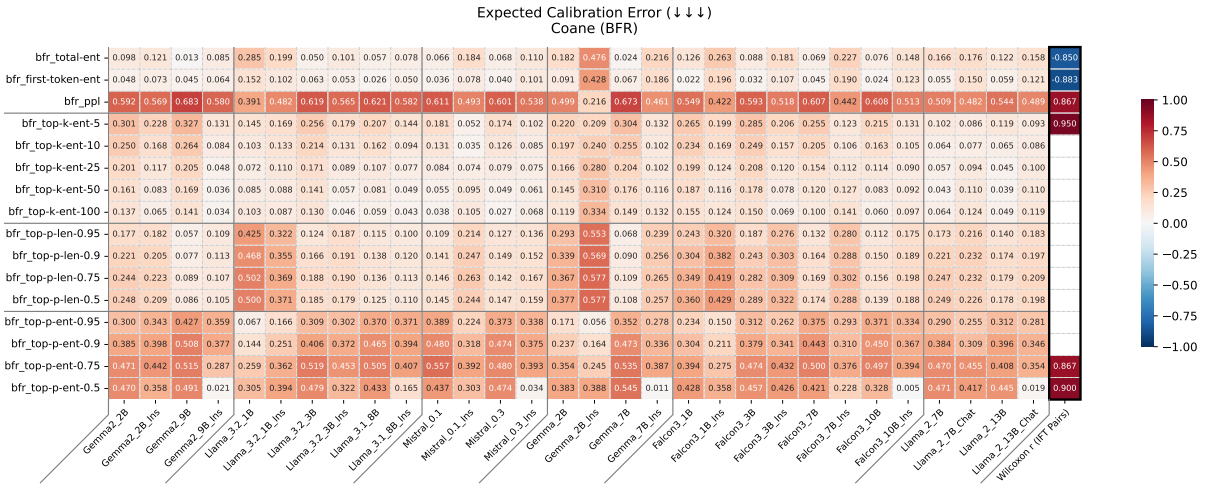

Figure 13: ECESweep results for the BFR version of the Coane dataset. Darker cells indicate higher calibration error and thus lower calibration. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models, with blank cells indicating insignificance. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

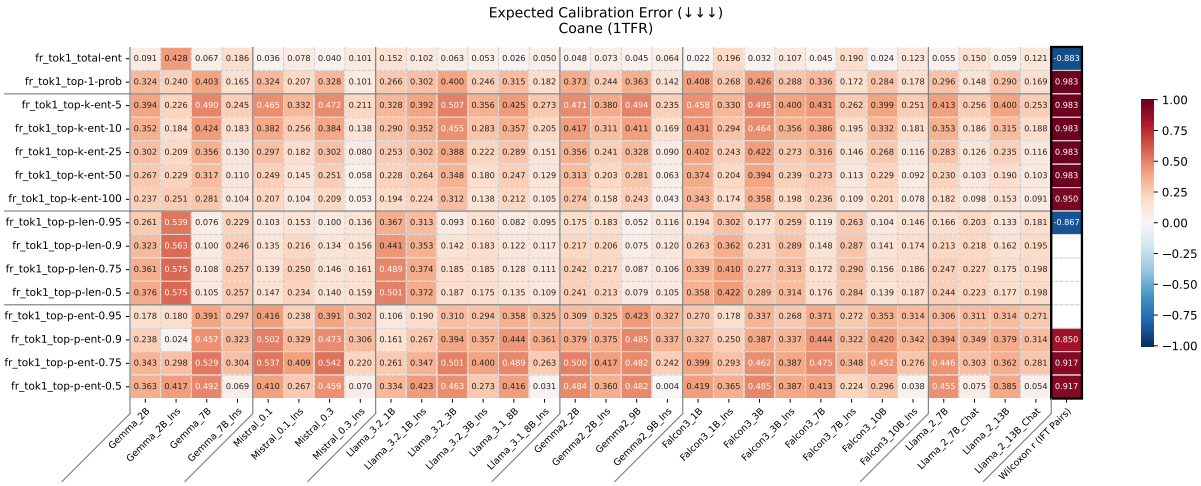

Figure 14: ECESweep results for the 1TFR version of the Coane dataset. Darker cells indicate higher calibration error and thus lower calibration. The rightmost column shows the Wilcoxon effect size between each pair of base vs instruct models, with blank cells indicating insignificance. Models are grouped by family, models are sorted within family by model size, and families are sorted by mean per-column maximum correlation

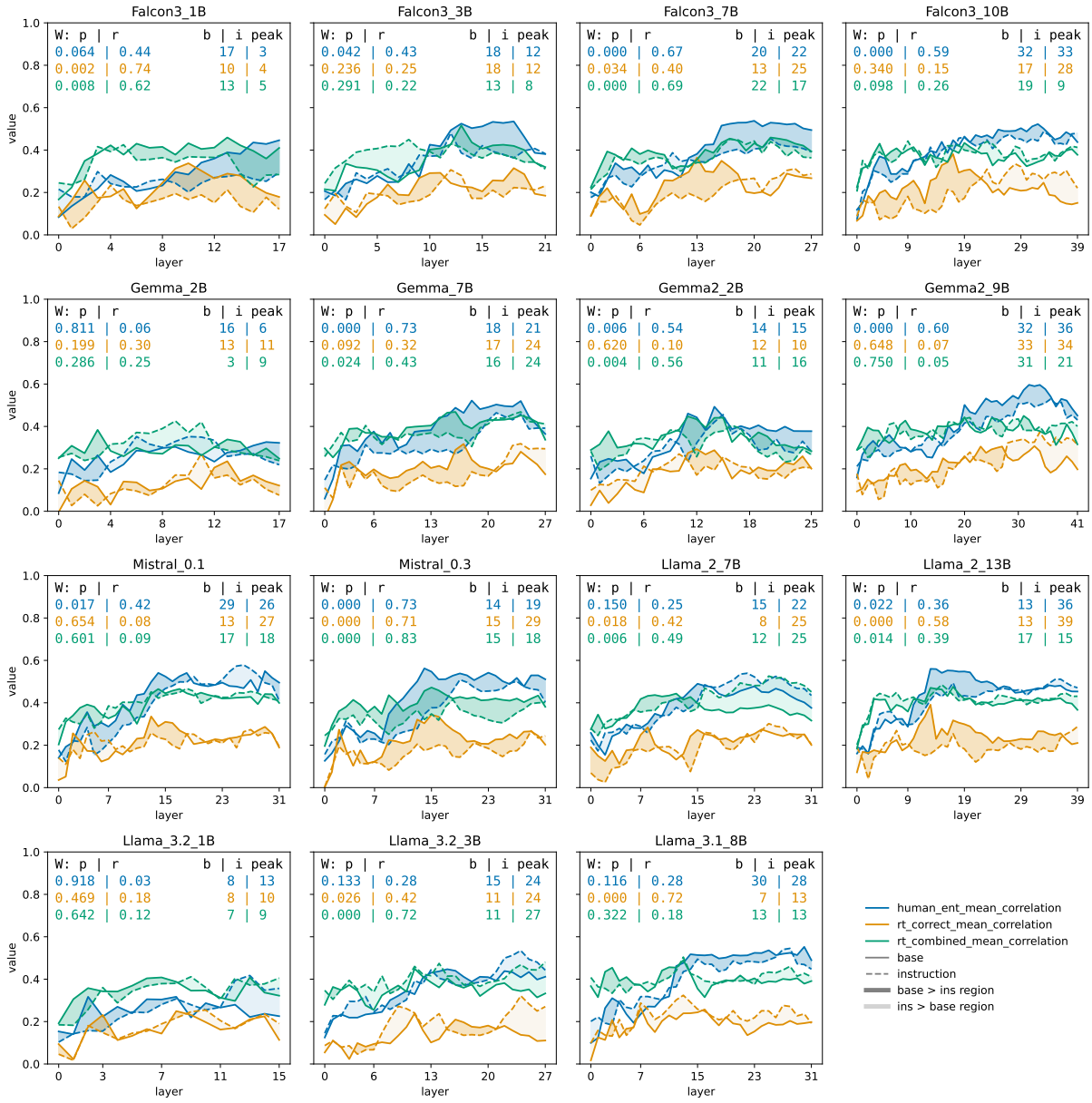

Figure 15: Per-layer model activation probing correlations. Colors indicate the human uncertainty type targeted during linear regression. Shaded regions indicate the gap between the base mode (solid line) and the instruct finetuned variant (dashed line). On-figure annotations in the top-left indicate the p-value and effect size for a Wilcoxon rank sum test between the base correlation and ins correlation across layers. Annotations in the top-right display the layer at which maximum human alignment was measured