On the Effect of Uncertainty on Layer-wise Inference Dynamics

Sunwoo Kim, Haneul Yoo, Alice Oh

{jaemo98, haneul.yoo}@kaist.ac.kr, alice.oh@kaist.edu

Paper

We show uncertainty is not significantly reflected in inference dynamics, which suggests existence of fundamental limitations in applying simple interpretability methods to measure uncertainty.

Research Questions

- LLMs encode uncertainty in hidden states, but it is not certain how this affects inference.
- Inference adaptive to uncertainty levels may arise as an emergent capability.

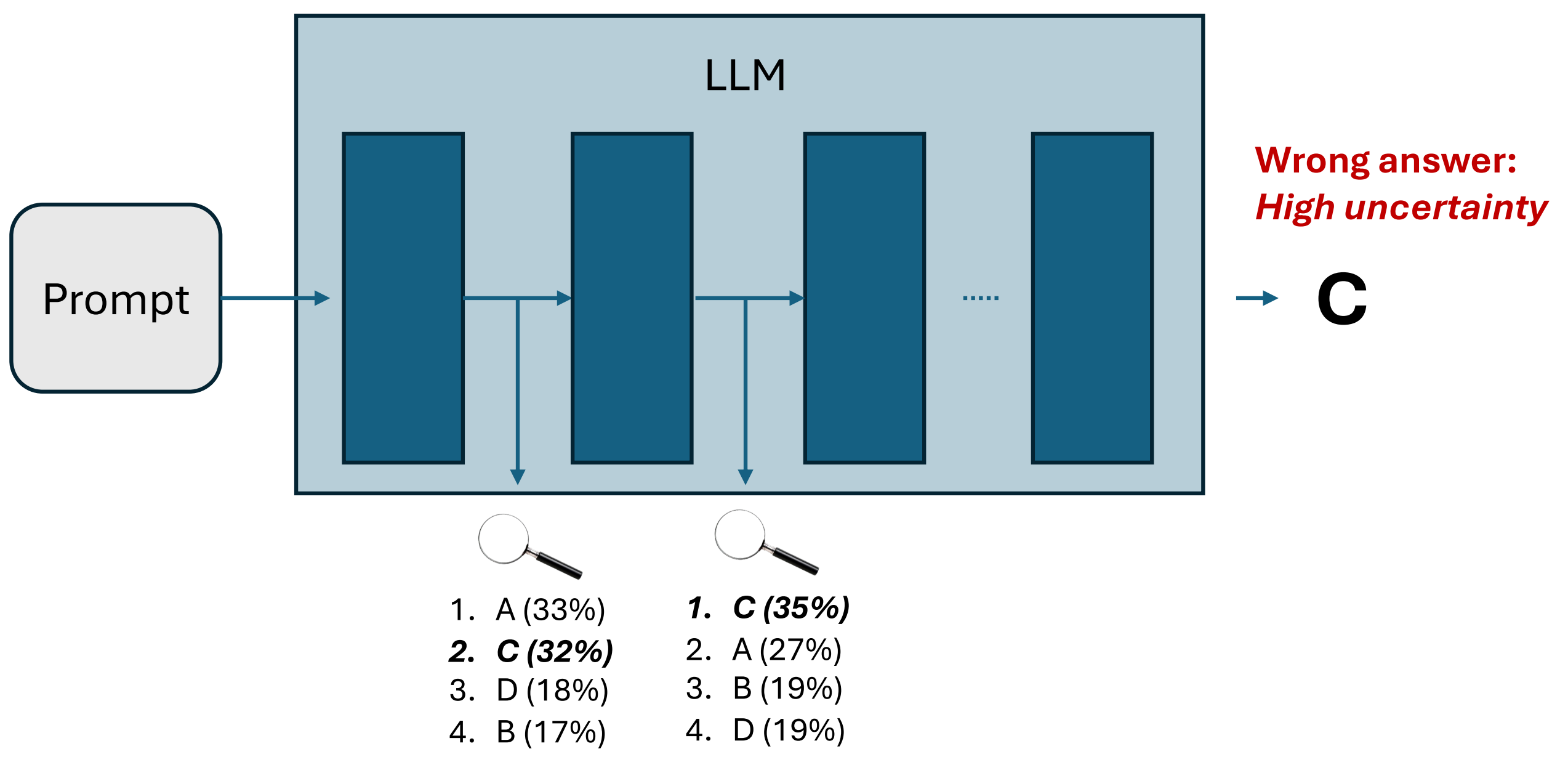
→ RQ1: How does uncertainty affect the inference dynamics of models?

→ RQ2: Does model competence affect the ability to adapt inference dynamics to uncertainty?

Methodology

- Extract intermediate logit distributions using Tuned Lens, a variant of Logit Lens.
- Observe inference dynamics in terms of ***intermediate token probabilities*** and ***prediction depth*** (PD, commitment layer after which the top prediction does not change).
- Model answers MCQs where ***questions answered wrongly are considered to induce high epistemic uncertainty***.
- We test on 5 models across 11 datasets.

Answer the question with a single letter like [A].
George wants to warm his hands quickly by rubbing them.
Which skin surface will produce the most heat?
A. dry palms
B. wet palms
C. palms covered with oil
D. palms covered with lotion
Answer: [



- Divergences in probability trajectories and PD, which reflects effective layers used, would imply ***greater adaptability in inference dynamics to uncertainty***.

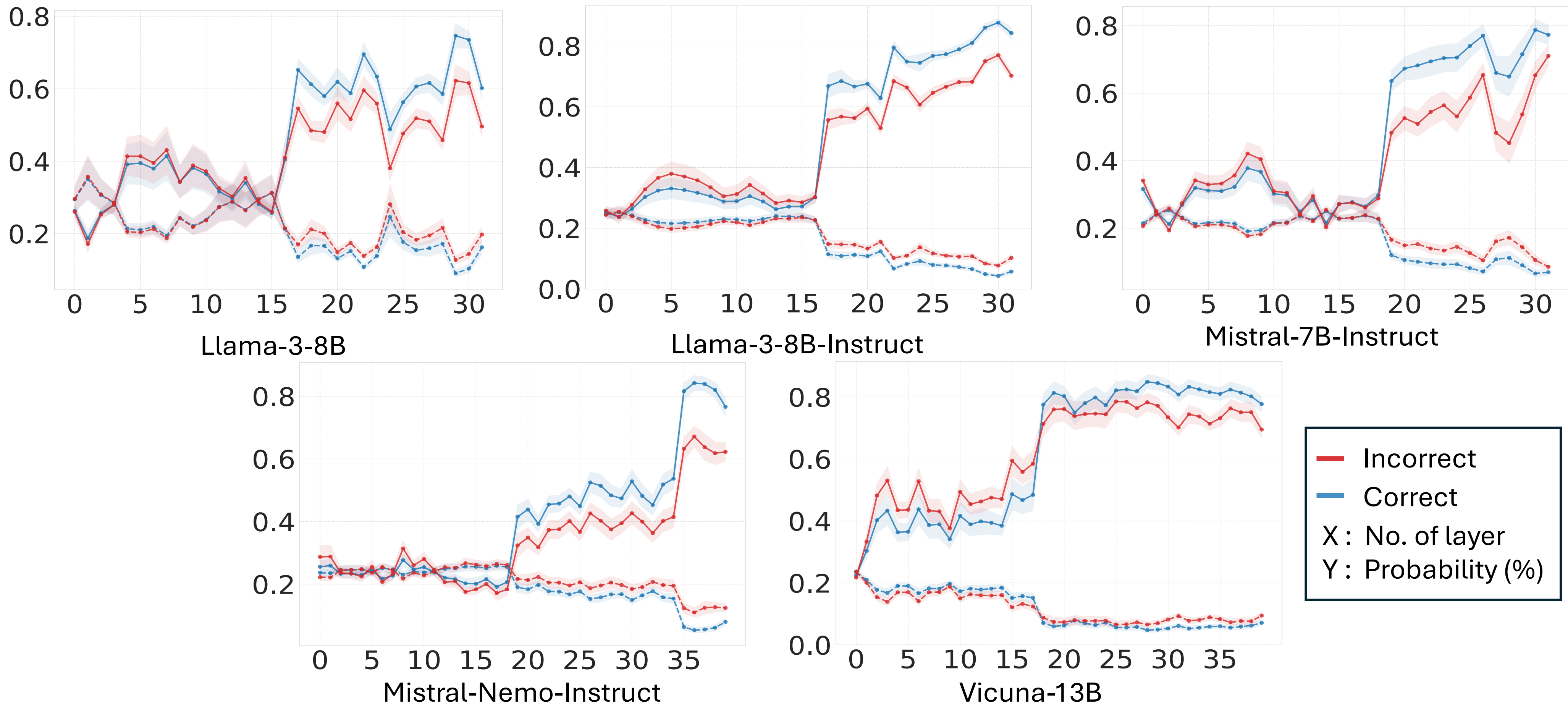
Conclusion

1. Uncertainty does not affect inference dynamics significantly.
2. Models abruptly decide on outputs at specific layers regardless of uncertainty.
3. Absolute effect of uncertainty is small, but greater task proficiency seems to cause greater adaptability to uncertainty.

→ Implies **limitations to applying simple interpretability methods to gauge model uncertainty**, especially on less competent models (e.g., hallucination detection).

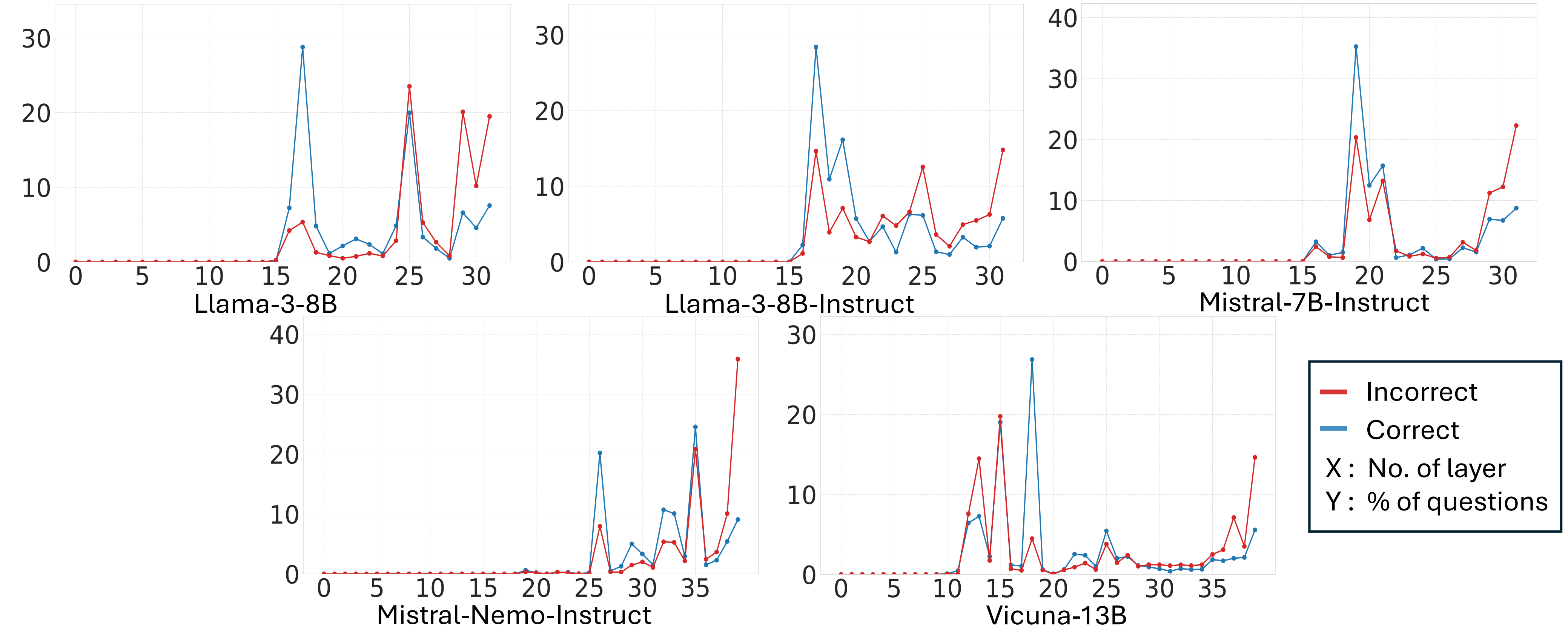
Results

Average probability trajectories for final output tokens across layers



Trajectories of correct and wrong outputs are aligned and model abruptly decides on final prediction at similar layers regardless of correctness.

Prediction depth distribution of questions across layers



Distributions are largely aligned, showing that model commits to answers at similar specific layers regardless of correctness.

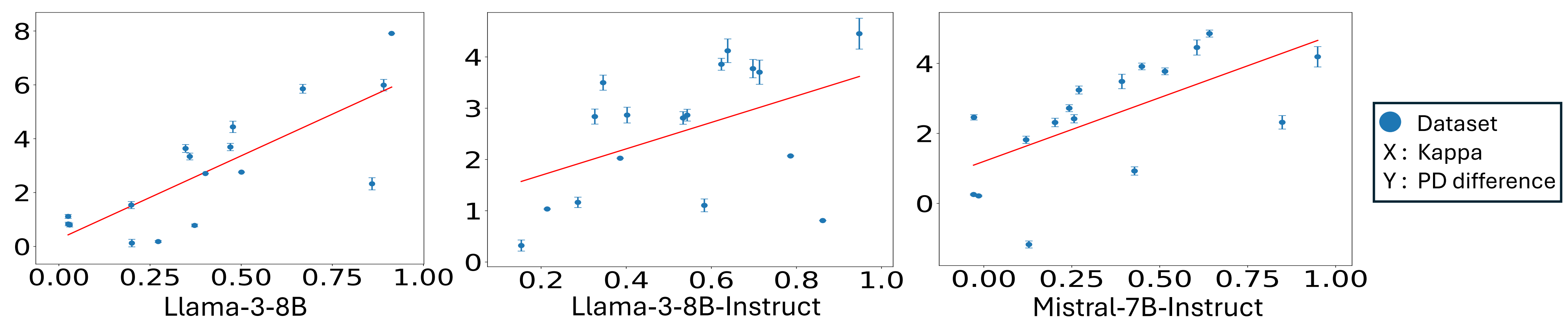
Pearson correlation between answer incorrectness and PD

Dataset	Llama-3-8B	Llama-3-8B-Instruct	Vicuna-13B	Mistral-7B-Instruct	Mistral-Nemo-Instruct
ANLI-R1	0.192*(0.015)	0.266*(0.013)	0.195*(0.014)	0.118*(0.014)	**0.428***(0.012)
ANLI-R2	0.174*(0.016)	**0.321***(0.013)	0.194*(0.014)	0.146*(0.014)	**0.464***(0.011)
ANLI-R3	0.238*(0.015)	0.258*(0.013)	0.215*(0.013)	0.284*(0.009)	0.240*(0.013)
ARC-Easy	**0.449***(0.011)	**0.351***(0.019)	0.069*(0.021)	**0.443***(0.017)	**0.332***(0.019)
ARC-Challenge	**0.373***(0.017)	**0.363***(0.017)	0.040(0.030)	**0.322***(0.018)	**0.357***(0.017)
BoolQ	0.061*(0.012)	0.159*(0.014)	0.058*(0.014)	-0.168*(0.014)	-0.083*(0.014)
BoolQ w/ context	0.175*(0.011)	0.132*(0.014)	0.233*(0.013)	0.107*(0.014)	-0.081*(0.014)
CommonsenseQA	**0.316***(0.011)	0.255*(0.009)	-0.045*(0.010)	0.274*(0.009)	0.258*(0.010)
HellaSwag	0.010(0.012)	0.047*(0.016)	0.188*(0.014)	0.261*(0.013)	**0.309***(0.013)
LogiQA	0.174*(0.014)	0.120*(0.011)	0.113*(0.014)	0.216*(0.011)	0.269*(0.015)
MMLU	**0.332***(0.013)	0.217*(0.008)	0.139*(0.014)	0.265*(0.009)	**0.364***(0.013)
QASC	0.287*(0.010)	0.252*(0.010)	0.129*(0.011)	0.213*(0.011)	0.248*(0.010)
QASC w/ context	**0.374***(0.010)	0.263*(0.010)	**0.408***(0.009)	0.298*(0.010)	0.164*(0.011)
QuAIL	**0.318***(0.013)	**0.337***(0.009)	0.276*(0.009)	**0.381***(0.009)	**0.334***(0.013)
RACE	**0.315***(0.013)	**0.327***(0.013)	0.247*(0.013)	**0.370***(0.009)	**0.396***(0.012)
SciQ	**0.480***(0.007)	0.213*(0.009)	0.123*(0.014)	**0.484***(0.008)	0.101*(0.014)
SciQ w/ context	0.217*(0.014)	0.083*(0.009)	0.192*(0.014)	0.286*(0.013)	0.065*(0.014)

$^*p < 0.05$

Greater values mean that incorrectness correlates with later commitment, which implies *greater adaptability in terms of layers used*. Values are small, meaning uncertainty does not affect inference dynamics significantly, but prevalence of positive correlation shows potential for adaptive inference.

Kappa (accuracy) vs. avg. PD difference for correct and incorrect answers on datasets



Average difference in PD between correct and incorrect outputs shows difference in layers used according to uncertainty. Greater differences imply greater adaptability to uncertainty. As task competence measured by dataset accuracy increases, adaptability increases.