Fine-Tuning Language Models to Know What They Know

Sangjun Park, Elliot Meyerson, Xin Qiu, Risto Miikkulainen

NEURAL INFORMATION PROCESSING SYSTEMS

Fine-Tuning Language Models to Know What They Know

Metacognition is hard to measure in LLMs: self-reports are confounded by response biases and heuristics such as question difficulty or a habit of answering "Yes". We build a bias-controlled framework around the d'type2 metric and propose ESMA, an evolution-strategy fine-tuning method that aligns an LLM's self-reports with the correctness of its answers. Robust generalization was validated across unseen datasets, languages, and newly acquired knowledge. Gradual parameter updates reveal that these improvements are driven by sparse parameters.

- **Dual-question protocol:** a direct question and an independent meta question ("Do you know the answer?") are jointly evaluated.
- d'_{type2} measures discrimination ability with avoiding response bias.
- **ESMA:** gradient-free fine-tuning method based on Evolution Strategies (ES) with a joint reward on correctness and alignment.

Measuring Metacognition in LLMs

A dual-questioning protocol is implemented to evaluate LLMs. For each data point, models have to answer: (1) direct questions to elicit factual answers, and (2) meta questions to evaluate self-knowledge.

Human metacognition is measured by having subjects evaluate their own task accuracy. We applies these psychological paradigms in LLMs.

$$d'_{\text{type2}} = \Phi^{-1}(\text{Hit Rate}) - \Phi^{-1}(\text{False Alarm Rate})$$

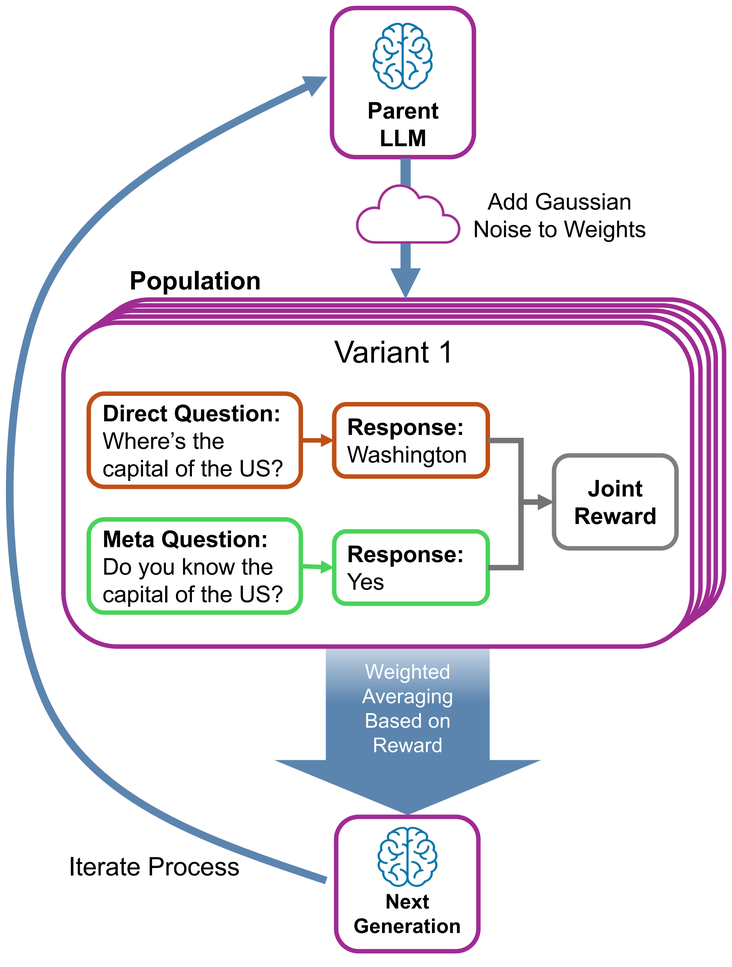
d'_{type2} measures the capacity to discriminate between accurate and inaccurate judgments by calculating the distance between the internal confidence distributions for each decision type.

Evaluation Result

Model	d'_{type2}	Raw Alignment	Accuracy
Qwen2.5 1.5B	0.20	53.30%	**42.86%**
Qwen2.5 1.5B ESMA	**0.93**	**68.86%**	41.86%
Qwen2.5 3B	0.29	62.70%	35.67%
Qwen2.5 3B ESMA	**1.02**	**69.59%**	**51.20%**
Qwen2.5 7B	0.64	61.65%	50.43%
Qwen2.5 7B ESMA	**0.94**	**69.86%**	**60.71%**

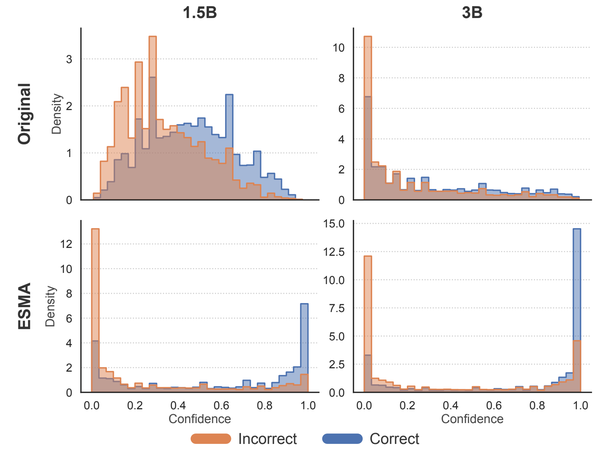
Across all models, ESMA substantially increases d'_{type2} and raw alignment metrics, showing that it improves metacognitive discrimination beyond a simple shift in the overall response strategy

ESMA: Evolution Strategy for Metacognitive Alignment



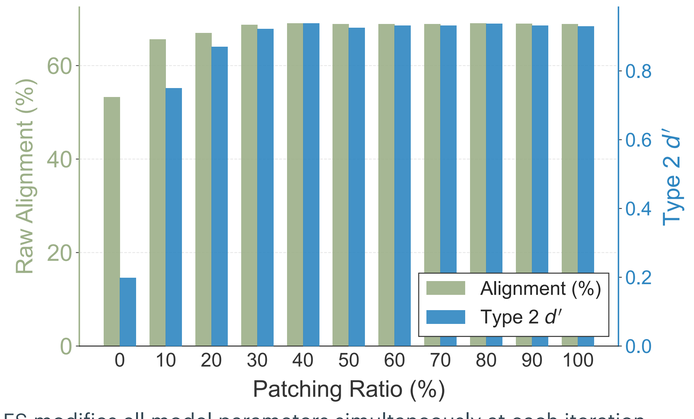
ESMA generates a candidate population by adding Gaussian noise to the model parameters. Each variant answers a direct and a meta question in independent passes, and a joint reward R(C, A) = C + A scores correctness (C) plus alignment (A) of the meta answer. The next generation LLM is a normalized reward-weighted average of the variants. Through iterative generations, this process progressively explores the parameter space for configurations that jointly optimize correctness and metacognitive alignment.

Confidence Distribution



The plots show the confidence for correct and incorrect direct answers. The original models have substantially overlapping distributions, indicating that correct and incorrect answers often receive similar confidence. After applying ESMA, confidence shifts toward 1.0 for correct answers and 0.0 for incorrect ones, reducing the overlap between the two distributions.

Sparse Weight Effect



ES modifies all model parameters simultaneously at each iteration. However, even when a modification improves fitness, it is not necessary that every individual parameter change contributes positively to the fitness. The plot shows that substantial gains were observed with patching only the top 10% of parameters, after which performance plateaued. This observation suggests that metacognitive improvements can be achieved by training a sparse subset of parameters.