UNCOVERING UNCERTAINTY IN TRANSFORMER INFERENCE

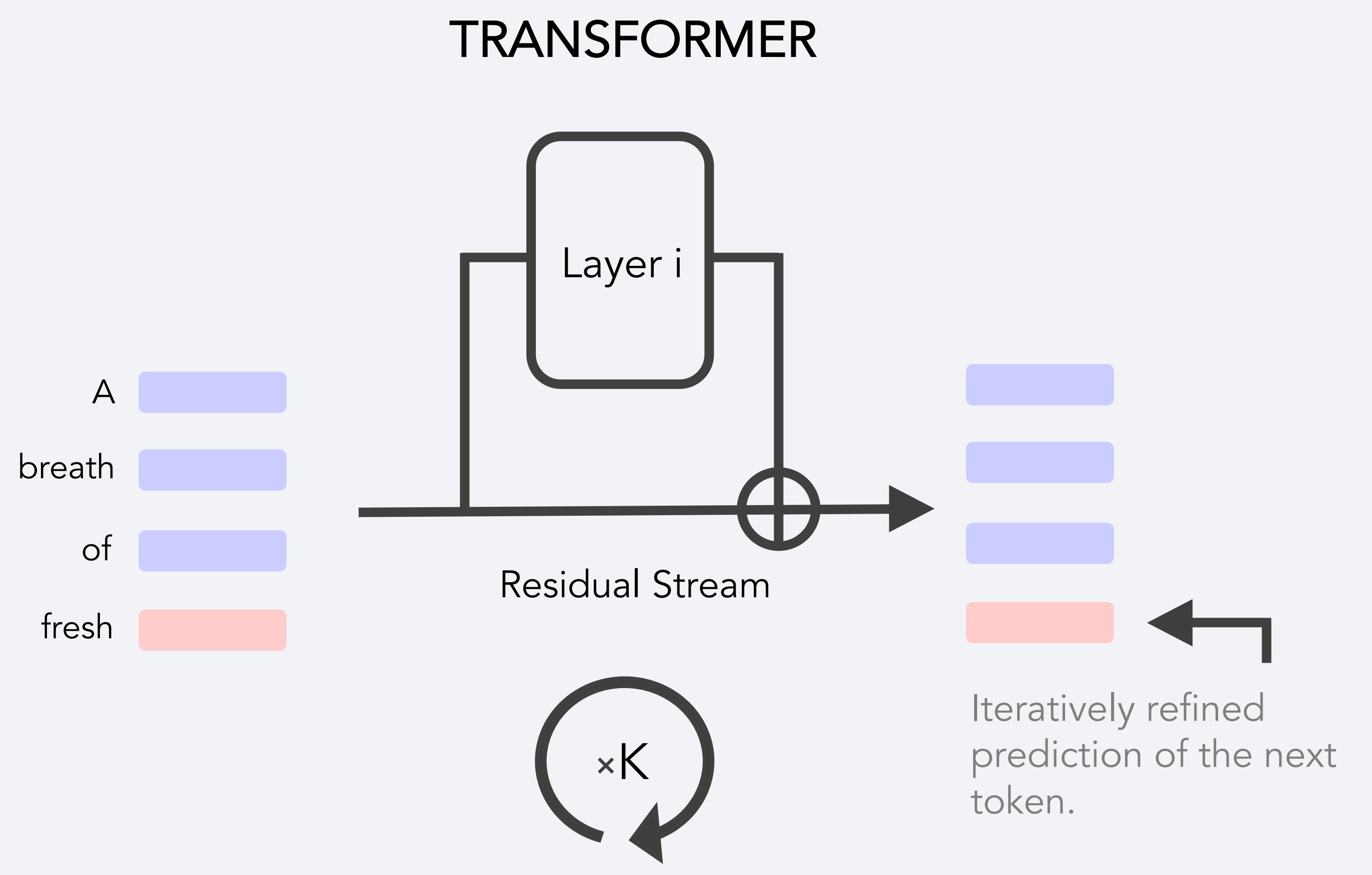
Greyson Brothers*, Willa Mannering, Amber Tien, John Winder

JOHNS HOPKINS APPLIED PHYSICS LABORATORY

NEURAL INFORMATION PROCESSING SYSTEMS

Background

Our research stems from the *Iterative Inference Hypothesis*: residual architectures iteratively refine predictions during inference. For a transformer applied to autoregressive language tasks, this looks like the following:


TRANSFORMER

Layer i

A

breath

of

fresh

Residual Stream

×K

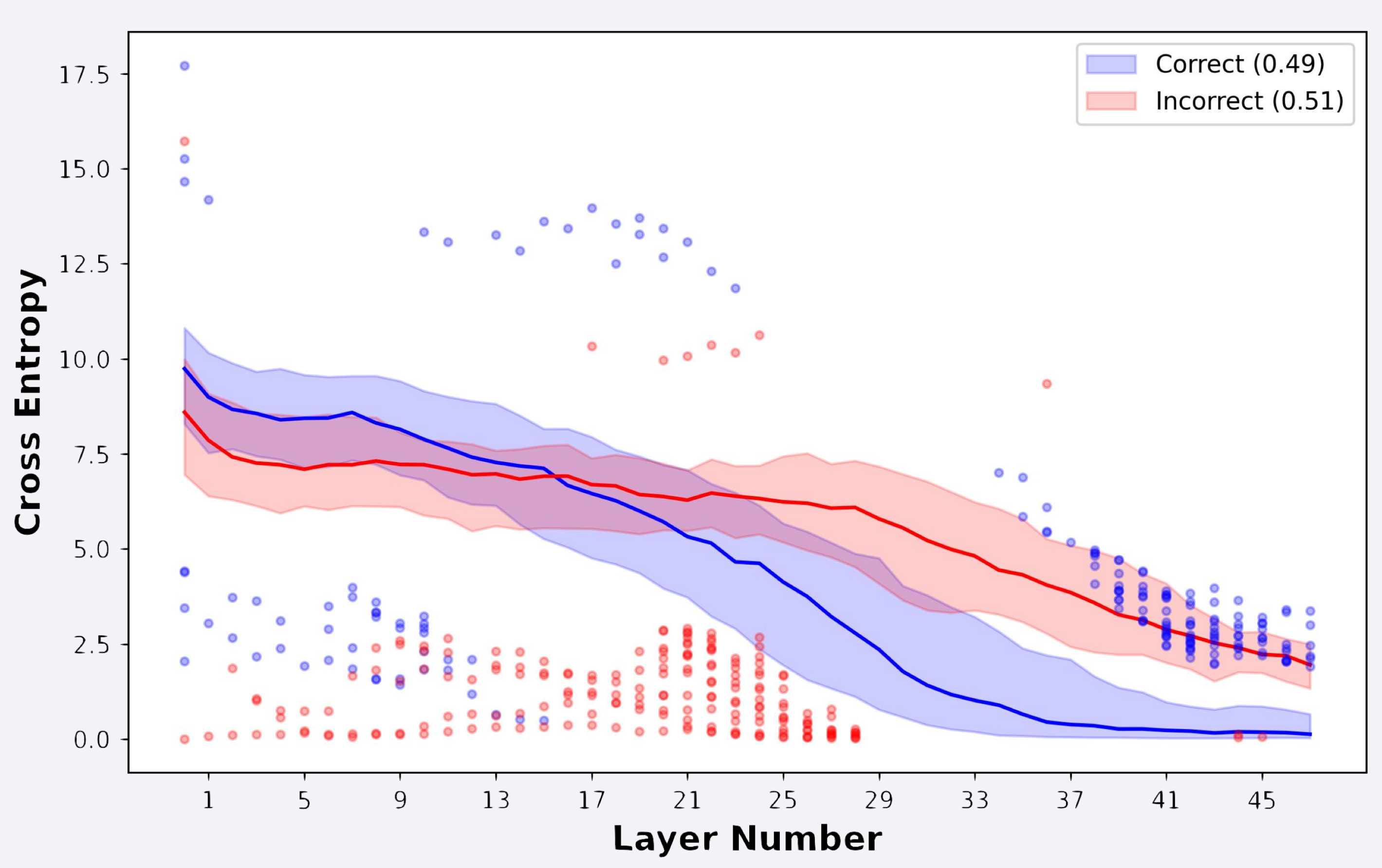
Iteratively refined prediction of the next token.

The nth embedding after each layer update can be passed through the unembed layer to obtain output probabilities over the vocabulary. This technique is called the logit lens.

Hypothesis

The more layers it takes to converge to a stable representation in the residual stream, the more uncertain the model is.

Convergence in Representation by Layer



Above we see how the distribution predicted by representations in the residual stream converges towards a stable output representation deeper in the model. We observe faster convergence when the model completed idioms correctly.

Dataset

EPIE Idiom Dataset

We chose an idiom dataset as it consisted of common single token completion prompts with a balanced range from heavily implied to extremely open ended.

Example idioms are provided below. We tasked GPT-2 XL with predicting the red word from the preceding context.

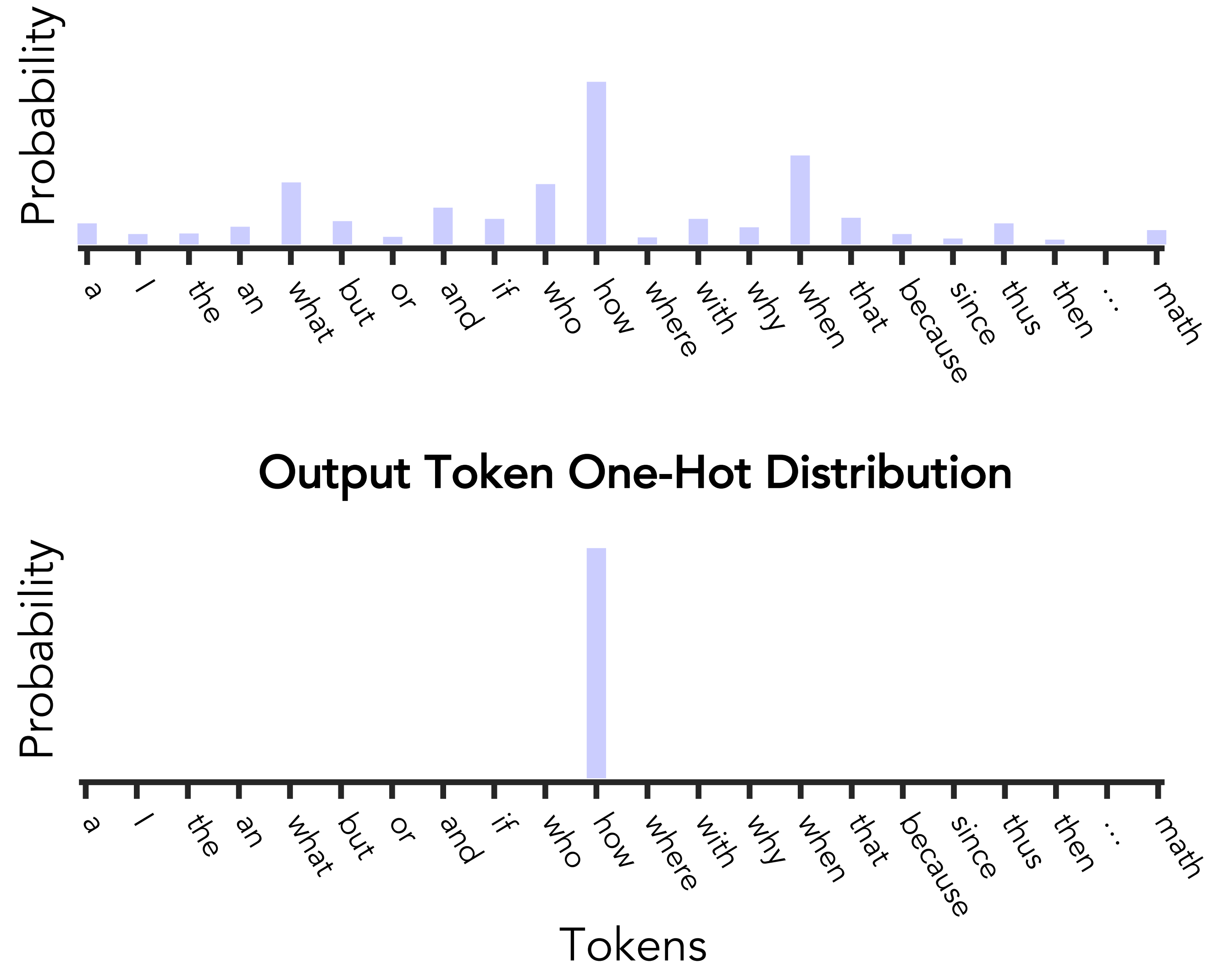
- "Great minds think *alike*"
- "Business as *usual*"
- "Every cloud has a silver *lining*"
- "On the *record*"
- "Behind closed *doors*"

Methods

We tracked the cross-entropy between the following two distributions after each layer update:

Prediction after the ith Layer

Output Token One-Hot Distribution

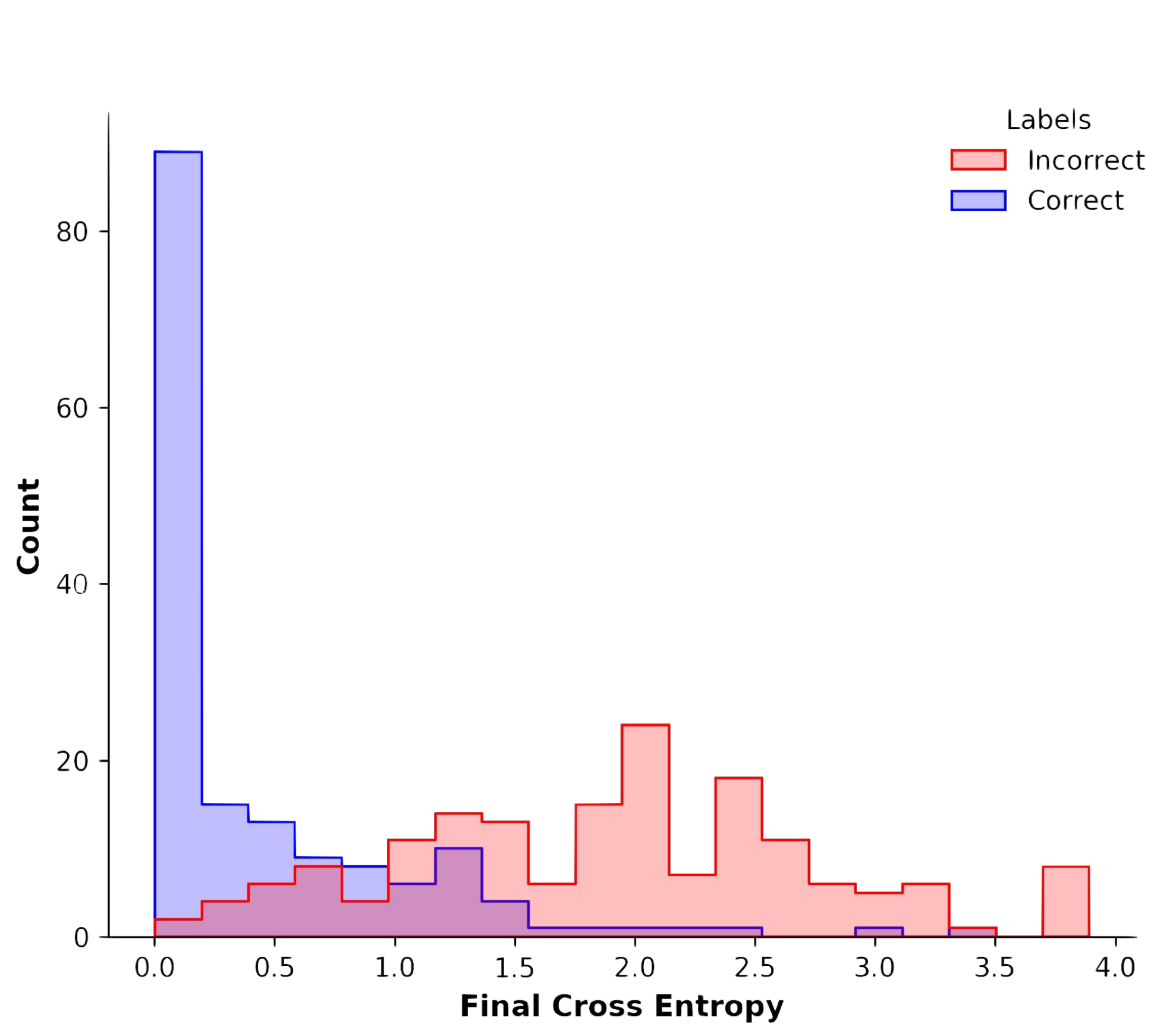


The latter distribution refers to the token that is actually sampled from the output probabilities of the final layer. This gives us a notion of how quickly and how closely the model's iteratively refined prediction converges to the final token it generates. Equivalent to the negative log likelihood of the output.

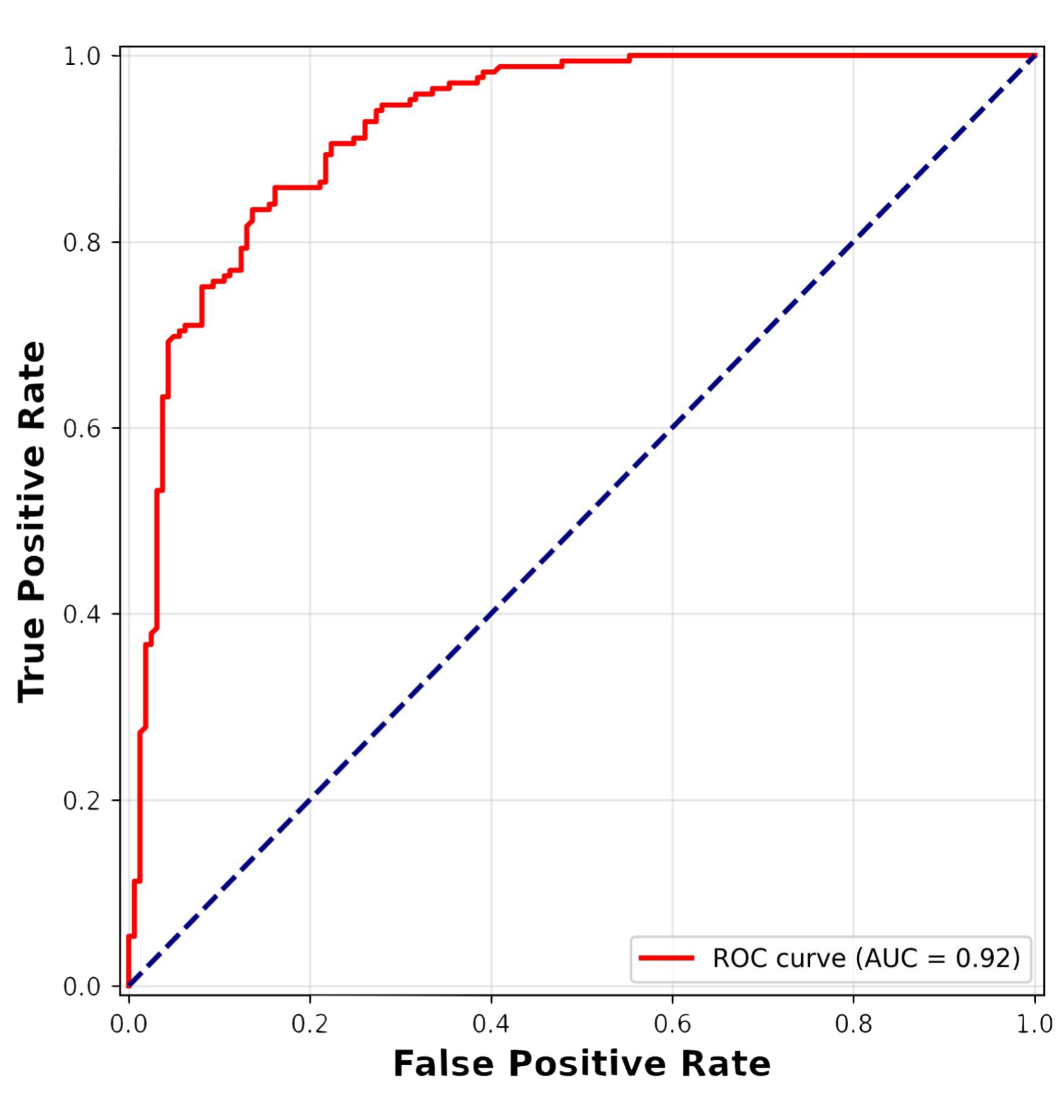
Findings

Residual Cross-Entropy at the final layer was highly discriminative of correct and incorrect idiom completion (left). We thus examine it's discriminative ability using an ROC curve (right).

Distribution of Cross-Entropy at the Final Layer



ROC Curve for Final Layer Cross-Entropy

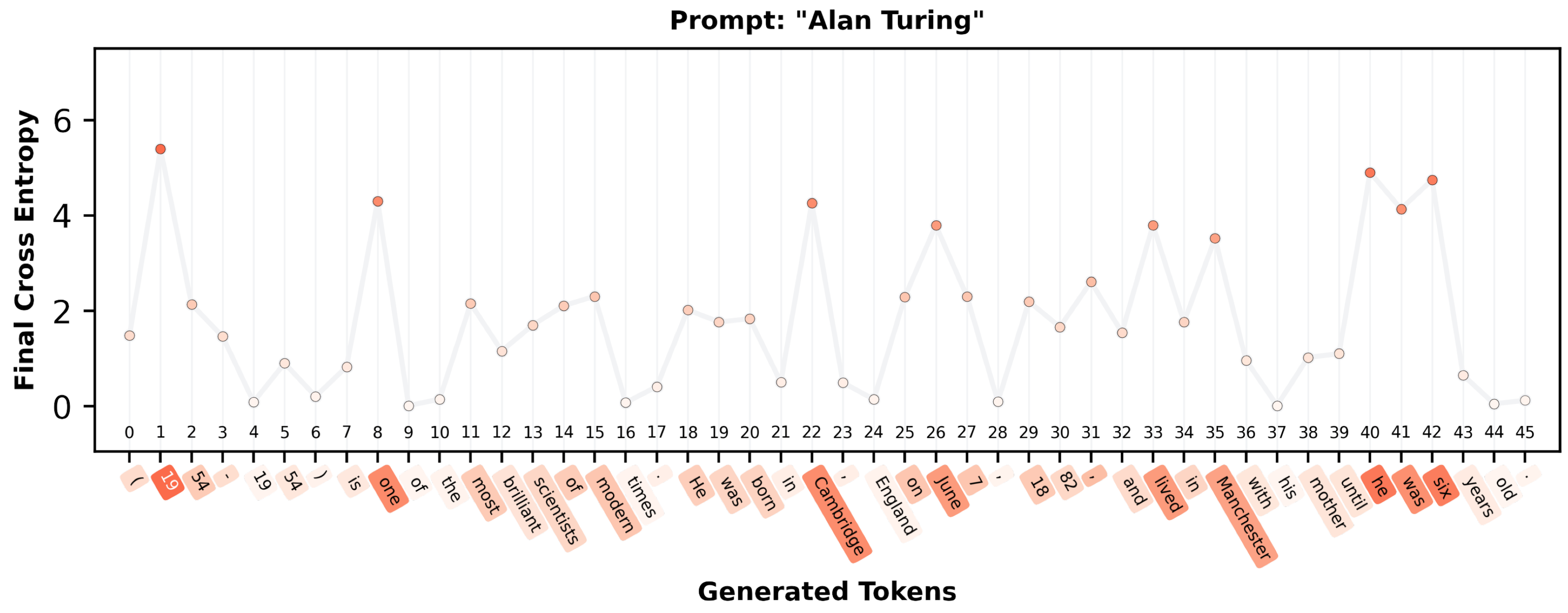


We observe an area under the curve of 0.9239, indicating that output cross-entropy is a strong predictor of correct vs incorrect generation on the idiom dataset.

Open Ended Example

Below we plot the final layer cross-entropy per token on an open-ended generation task. As can be observed, the measure captures a notion of uncertainty of the model given the prompt. For reference, Alan Turing was born in 1912 in London, England, attended King's College in Cambridge, and died on June 7, 1954.

Prompt: "Alan Turing"



* greyson.brothers@jhuapl.edu