

supermodel labs

Counting Tokens Before They Hatch

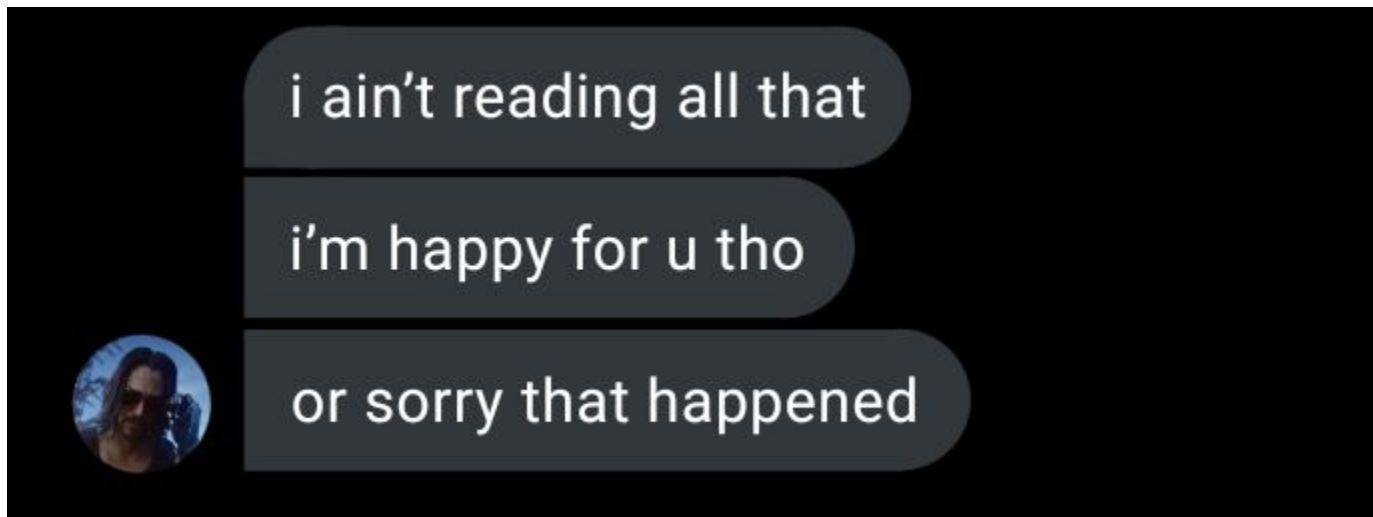
Predicting Agent Costs
Before Execution

Kirah Sapong

AGNTCon + MCPCon North America 2026

supermodellabs.com

- 1. What's your favourite colour?**
- 2. What did you do yesterday?**



supermodel labs

The terms we need

supermodellabs.com

01 Token

A model-specific unit of text

02 Embedding

A numerical representation of content

03 Regression

Predicting a quantity (numerical value)

Prompt: “Plan a complete trip from San Francisco to Bali...”

Chatbot

4,879

tokens

Agent

201,009 (41.2×)

tokens

**Can prompt context alone carry
enough predictive signal about
output size?**

The objective

Prompt

What is the capital of Japan?

Response

The capital of Japan is Tokyo (東京). However, the formal name of...

Target

87

supermodel labs

The starting point

supermodellabs.com

- 01 **Model** Zephyr-7B-Beta by HuggingFace
- 02 **Data** 304,246 prompt / response pairs
- 03 **Target** Token counts of final responses

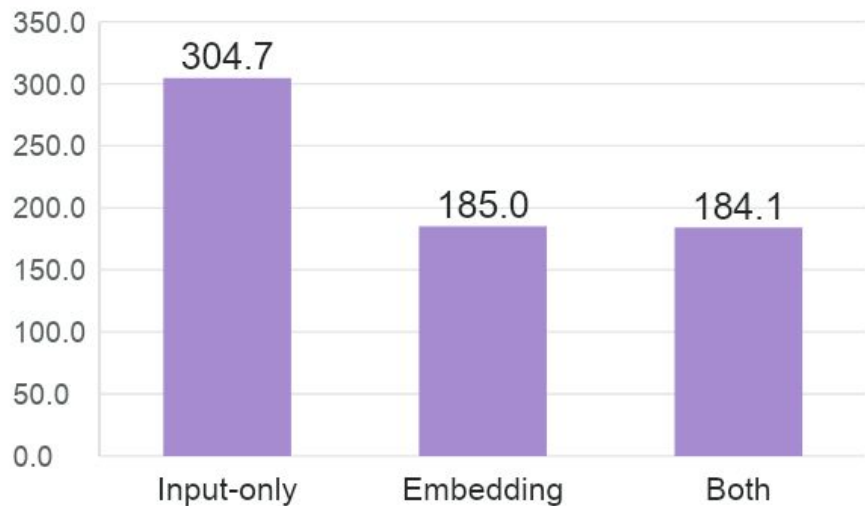
Three feature configurations

Configuration	Predictive features
Input-only	Input token count
Embedding-only	Semantic prompt embedding
Full model	Embedding + input token count

Embeddings reduce prediction error

↓ 40%

Lower mean absolute error



Mean absolute error in tokens. Lower is better.

Embeddings strengthen rank correlation

Spearman rank correlation improves with semantic features.

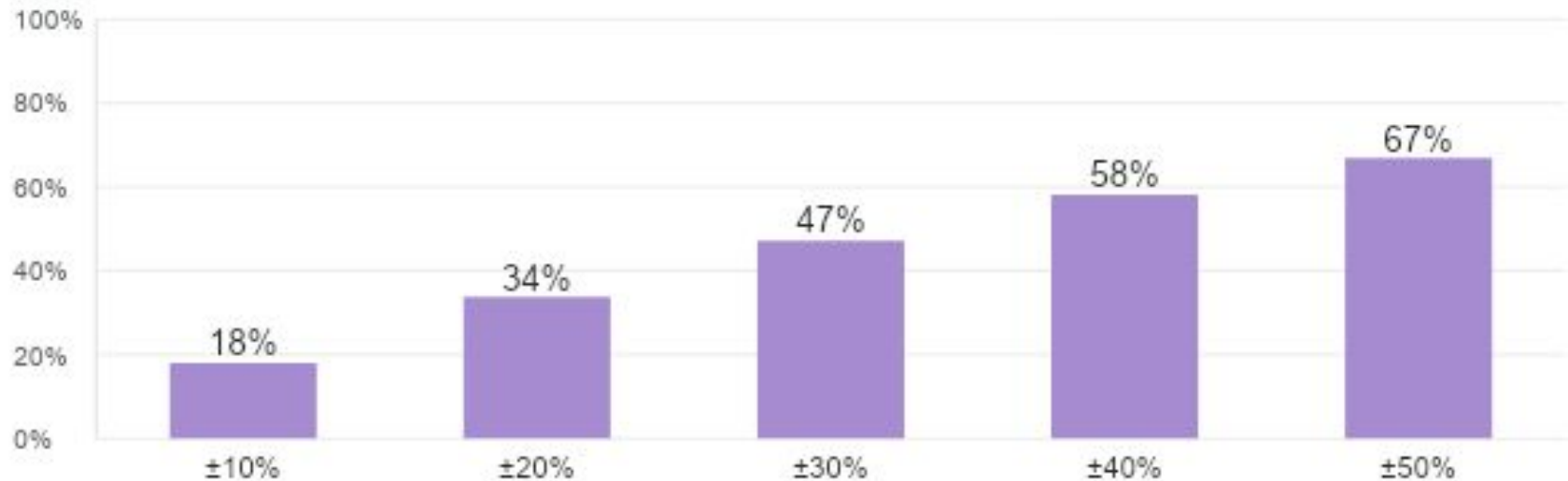
0.26

Input-only

0.77

↑ Full model

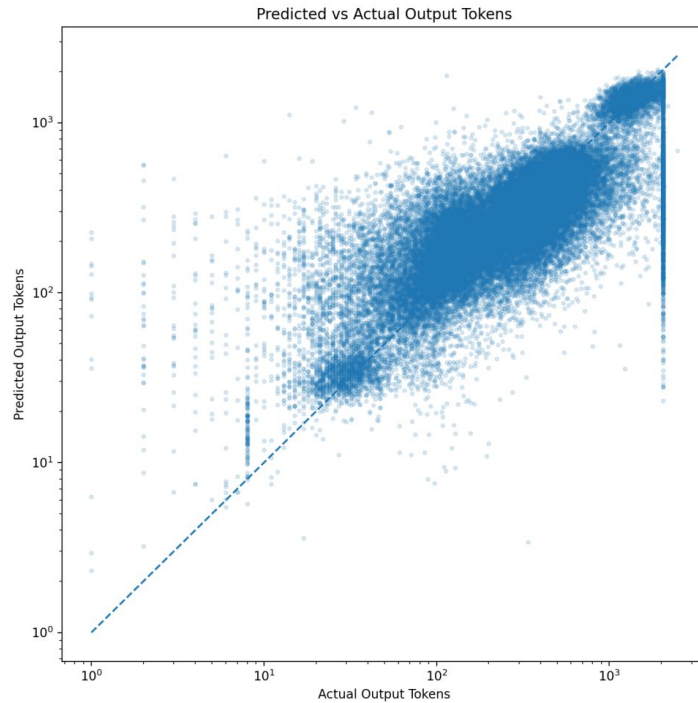
Half of predictions land within 30%



Share of full-model predictions inside each relative-error tolerance

The long tail creates budget risk

Long outputs often exceed the estimate



Open-ended vs. Factual

Prompt	Predicted	Actual
Tell me about AI.	215.8	553
Write a Python implementation of a trie with insert, search, and delete.	759.4	855

Agent forecasting has two uncertainties

1. How much will this call generate?
2. How many calls will the task require?

supermodel labs

Budget control

supermodellabs.com

- 01 **Estimate**
Known input, predicted output size
- 02 **Route**
Capability and remaining budget
- 03 **Execute**
Record actual usage after execution
- 04 **Reconcile**
Update the next decision

supermodel labs

Future work

supermodellabs.com

01 More model structures

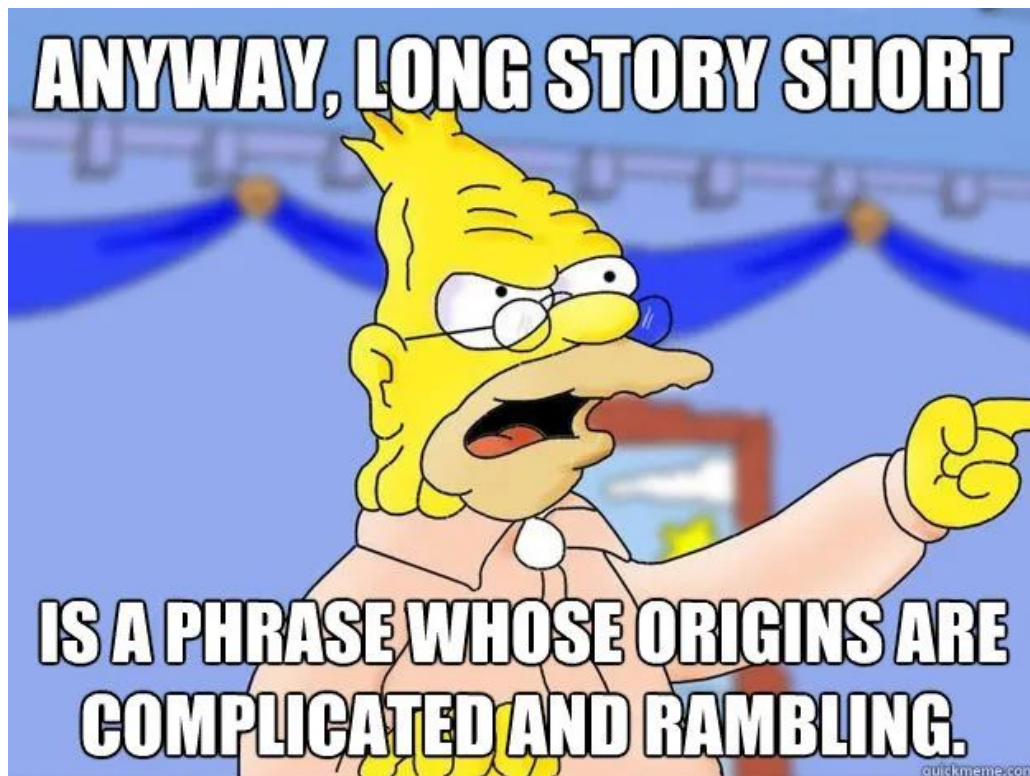
Decision trees, transformers

02 More conditions

Model family, hyperparameters

03 Stronger evaluation

Held-out models, longer contexts

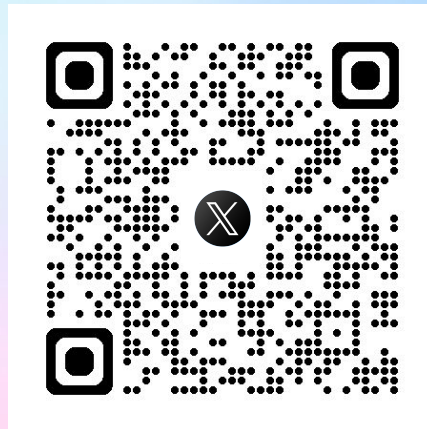


supermodel labs

Thank you

team@supermodellabs.com

supermodellabs.com



 @kirahsapong