QCD* in Language Models: What do they really know about QCD*?

Antonin Sulc[1] (main author) Patrick L.S. Connor[2] (speaker)

[1]LBNL

[2]CERN

10 July 2025

* or HEP in general

The Rise of LLMs in Science

- Large Language Models (LLMs) have shown remarkable capabilities in processing and generating human-like text.

  $\longrightarrow$ Some LLMs like Llama, Qwen, and Gemma are freely available ("open-weight models").
- LLMs are trained on publicly available texts.

  $\longrightarrow$ This includes a significant amount of scientific literature, textbooks, and articles.

- *How well do open-weight models actually understand HEP?*
- *How well do they relate concepts in HEP?*
- *How can they help us in doing science?*

The Rise of LLMs in Science

- Large Language Models (LLMs) have shown remarkable capabilities in processing and generating human-like text.

  $\longrightarrow$ Some LLMs like Llama, Qwen, and Gemma are freely available ("open-weight models").
- LLMs are trained on publicly available texts.

  $\longrightarrow$ This includes a significant amount of scientific literature, textbooks, and articles.

- *How well do open-weight models actually understand HEP?*
- *How well do they relate concepts in HEP?*
- *How can they help us in doing science?*

What is an LLM? At its core, a prediction engine

- LLMs are sophisticated next-word predictors.
- Given an input sequence, the model calculates a probability distribution for the next "token" (a word or piece of a word)

  $\longrightarrow$ This is analogous to predicting a system's next state based on its history.

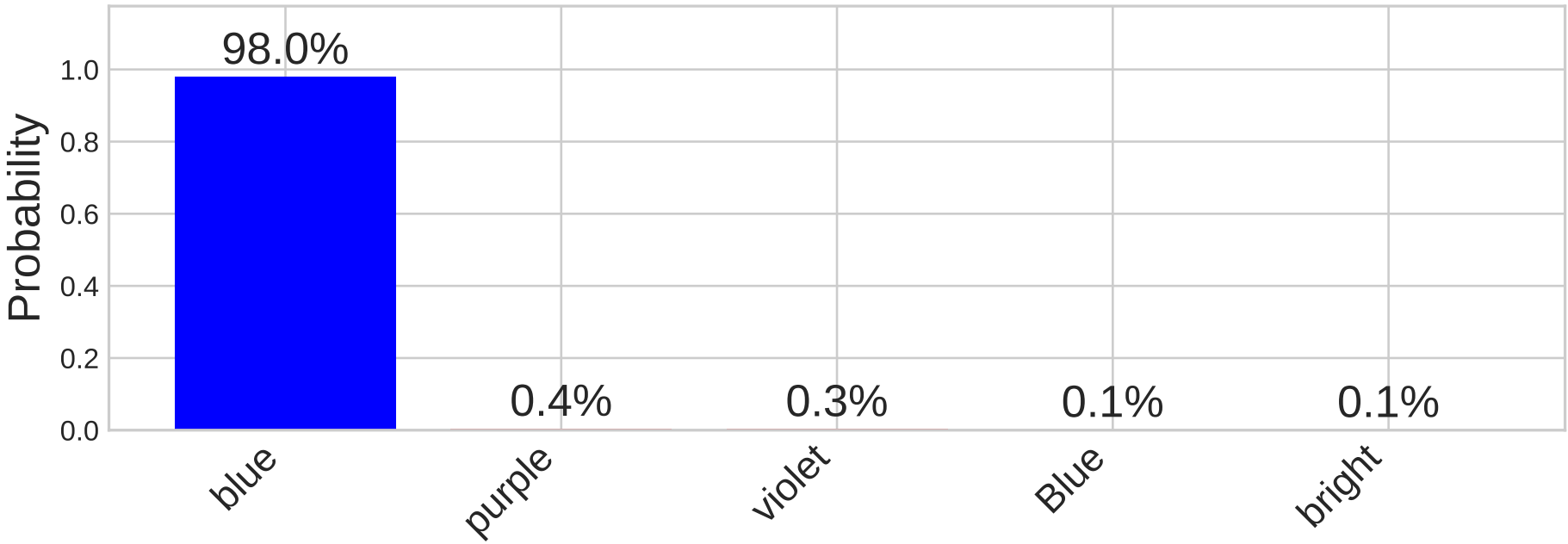
"Roses are red, violets are..."

What is an LLM? At its core, a prediction engine

- LLMs are sophisticated next-word predictors.
- Given an input sequence, the model calculates a probability distribution for the next "token" (a word or piece of a word)

  $\longrightarrow$ This is analogous to predicting a system's next state based on its history.

"Roses are red, violets are..."



Metric: measuring knowledge with perplexity

Perplexity (PPL) measures a model's "surprise" at a given text

$\longrightarrow$ lower perplexity means the model predicted the text more accurately.

It's the exponential of the average negative log-likelihood per token (w_i):

$$\mathrm{PPL}(W) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})\right)$$

Example: By measuring PPL, we can test the model's knowledge of a scientific assertion.

- **Low perplexity = good prediction:**
  "The strong force is mediated by gluons."

- **High perplexity = poor prediction:**
  "The strong force is mediated by photons."

Metric: measuring knowledge with perplexity

Perplexity (PPL) measures a model's "surprise" at a given text

$\longrightarrow$ lower perplexity means the model predicted the text more accurately.

It's the exponential of the average negative log-likelihood per token (w_i):

$$\mathrm{PPL}(W) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})\right)$$

Example: By measuring PPL, we can test the model's knowledge of a scientific assertion.

- **Low perplexity = good prediction:**
  "The strong force is mediated by gluons."
- **High perplexity = poor prediction:**
  "The strong force is mediated by photons."

Metric: measuring knowledge with perplexity

Perplexity (PPL) measures a model's "surprise" at a given text

$\longrightarrow$ lower perplexity means the model predicted the text more accurately.

It's the exponential of the average negative log-likelihood per token (w_i):

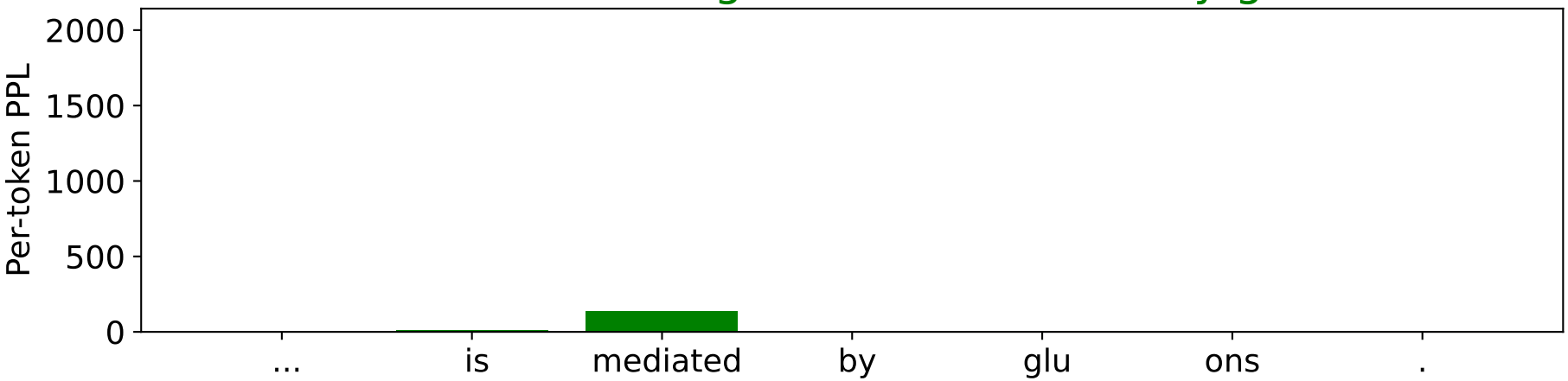
$$\mathrm{PPL}(W) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})\right)$$

Example: By measuring PPL, we can test the model's knowledge of a scientific assertion.

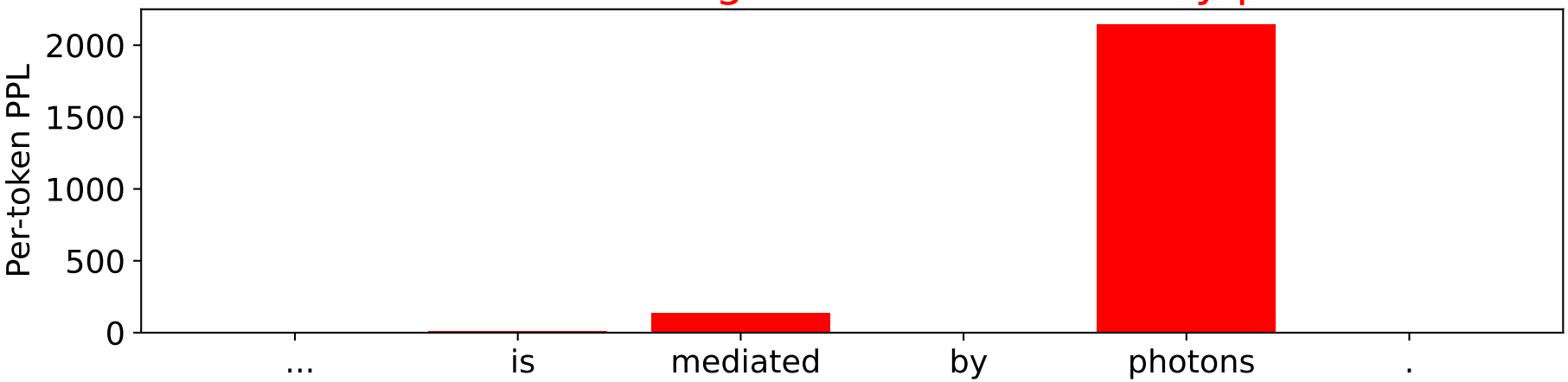
- **Low perplexity = good prediction:**
  "The strong force is mediated by gluons."
- **High perplexity = poor prediction:**
  "The strong force is mediated by photons."

Example of Perplexity

Correct: The strong force is mediated by gluons.



Incorrect: The strong force is mediated by photons.



Probing a fundamental constant: the value of $\alpha_s(M_Z)$

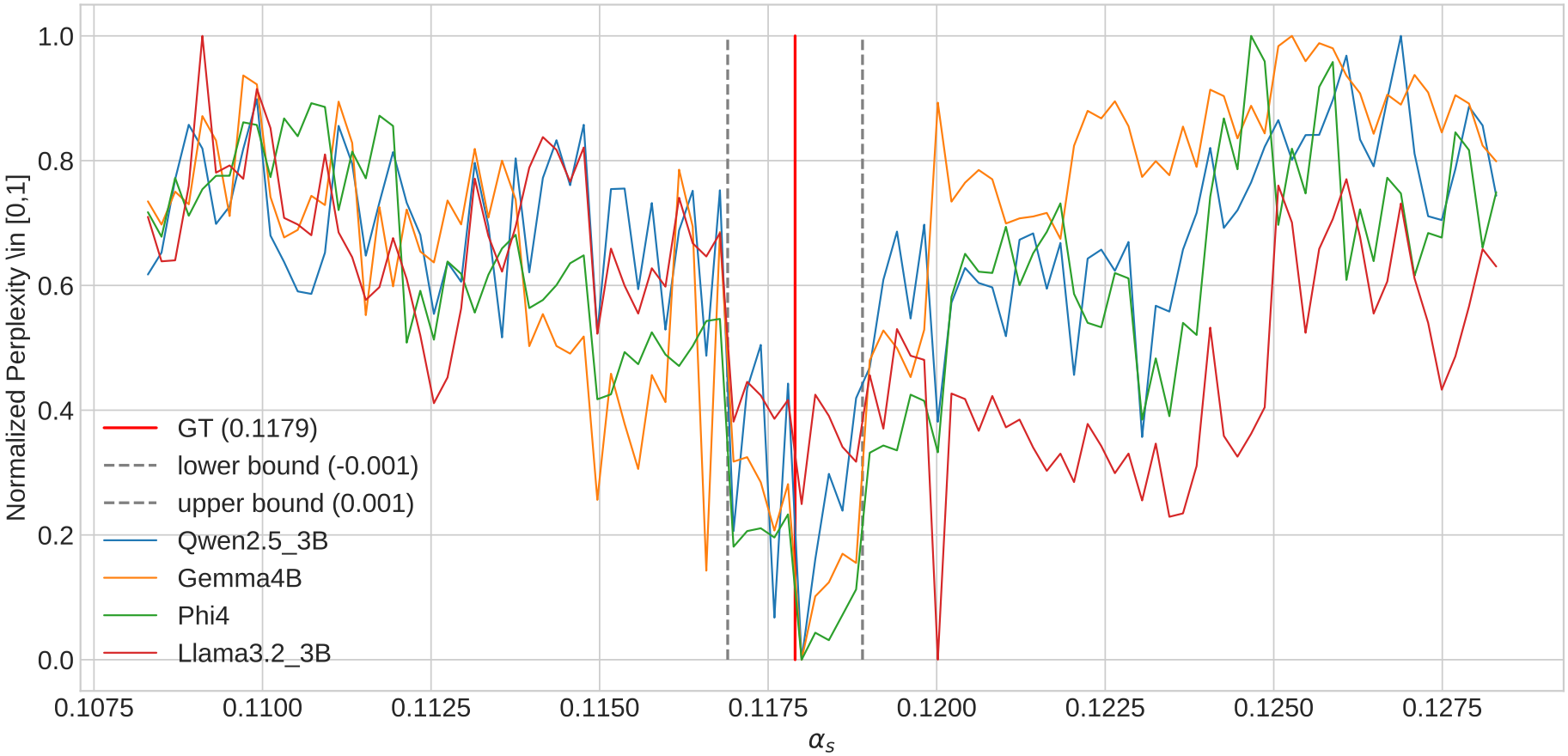
Question: how do models deal with numerical values?
Test: we feed the model prompts where only the value of the strong coupling constant, α_s, changes.

"Strong coupling constant α_s is $\{value\}$"

Probing a fundamental constant: the value of $\alpha_s(M_Z)$

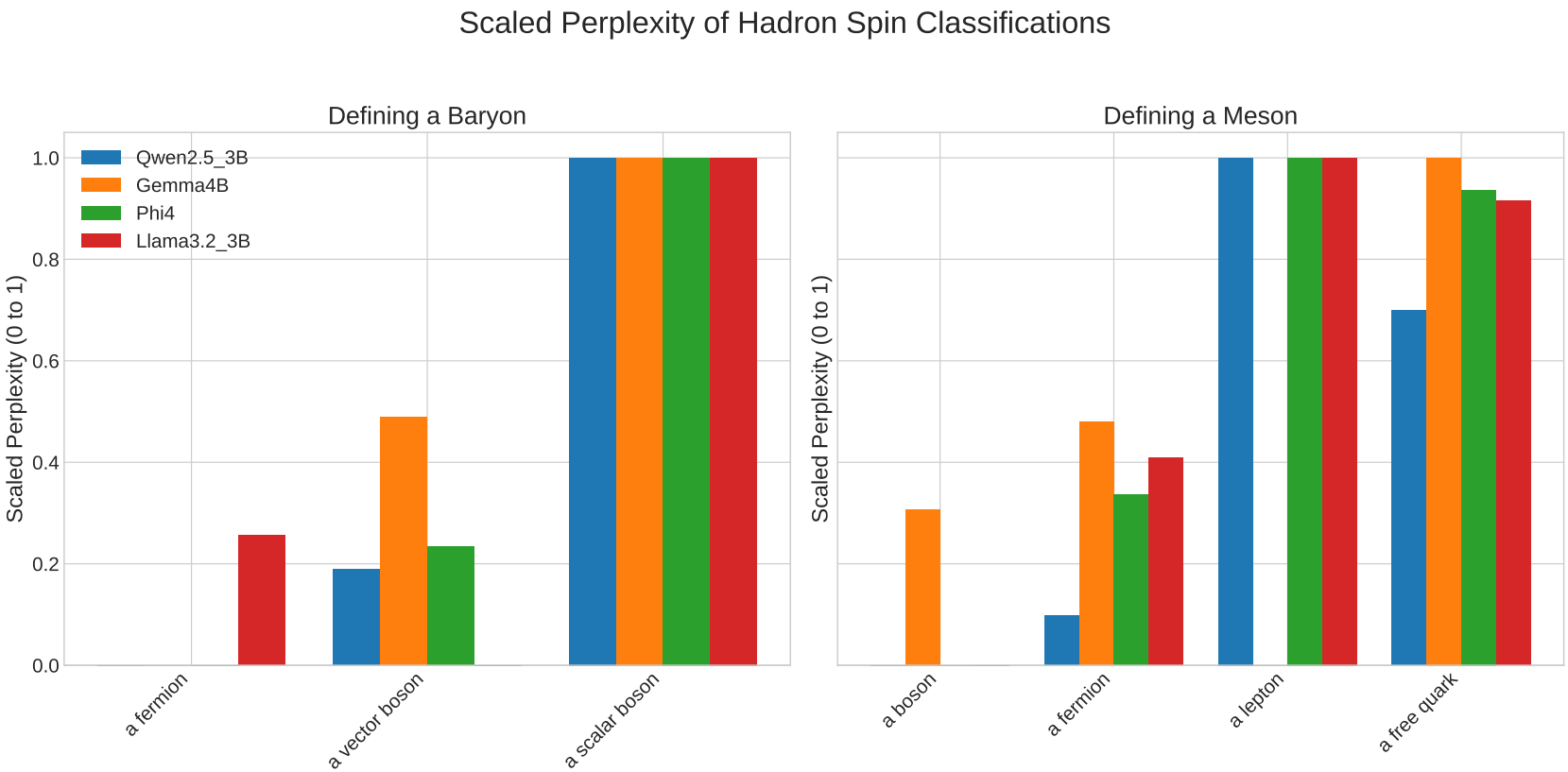
Question: how do models deal with numerical values?
Test: we feed the model prompts where only the value of the strong coupling constant, α_s, changes.

"Strong coupling constant α_s is {value}"



Classifying hadrons by spin: fermions vs. bosons

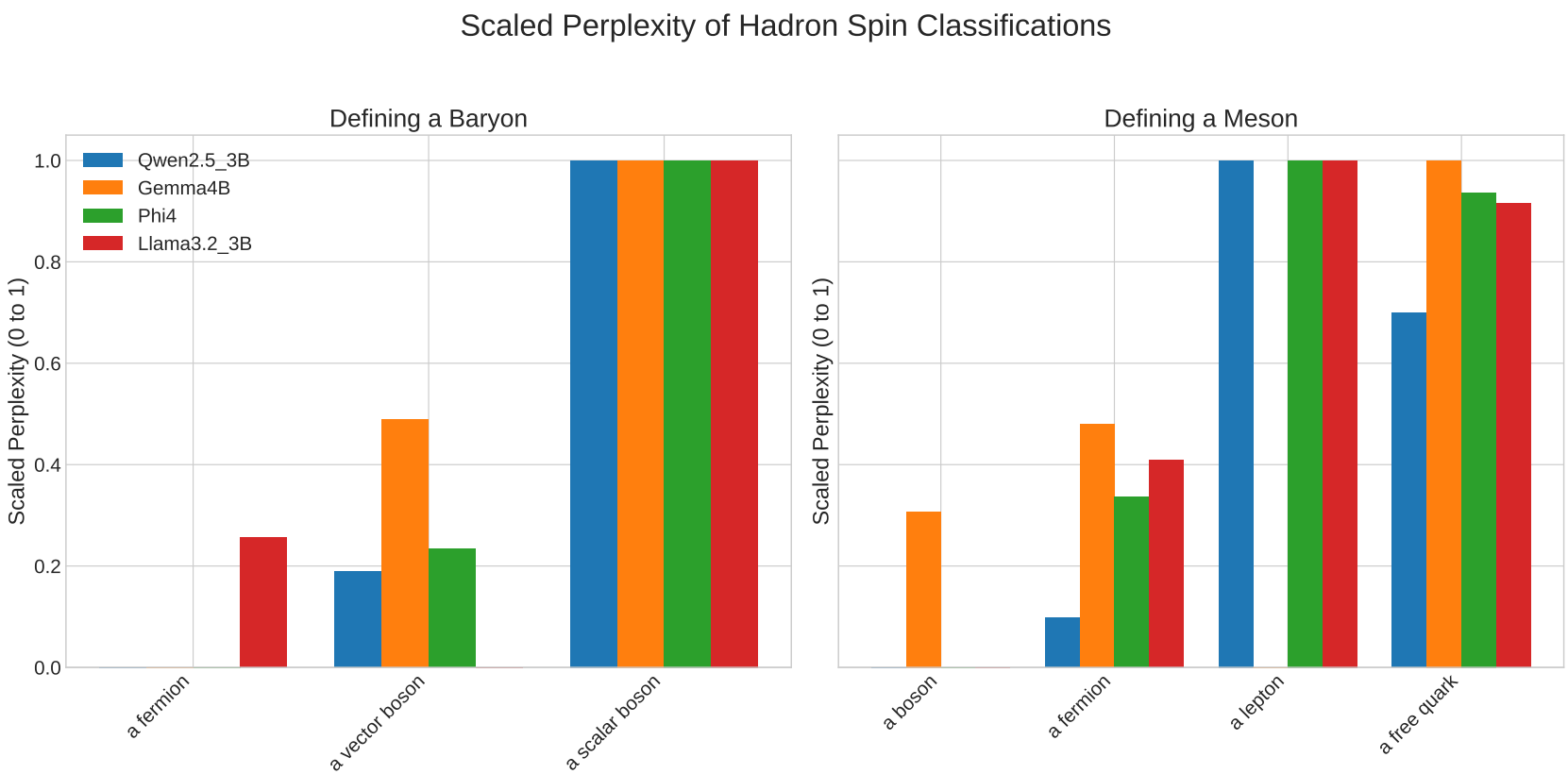
"Based on its total spin, a baryon is classified as {classification}."



$\longrightarrow$ Llamma3_2B and Gemma4B make small mistakes

Classifying hadrons by spin: fermions vs. bosons

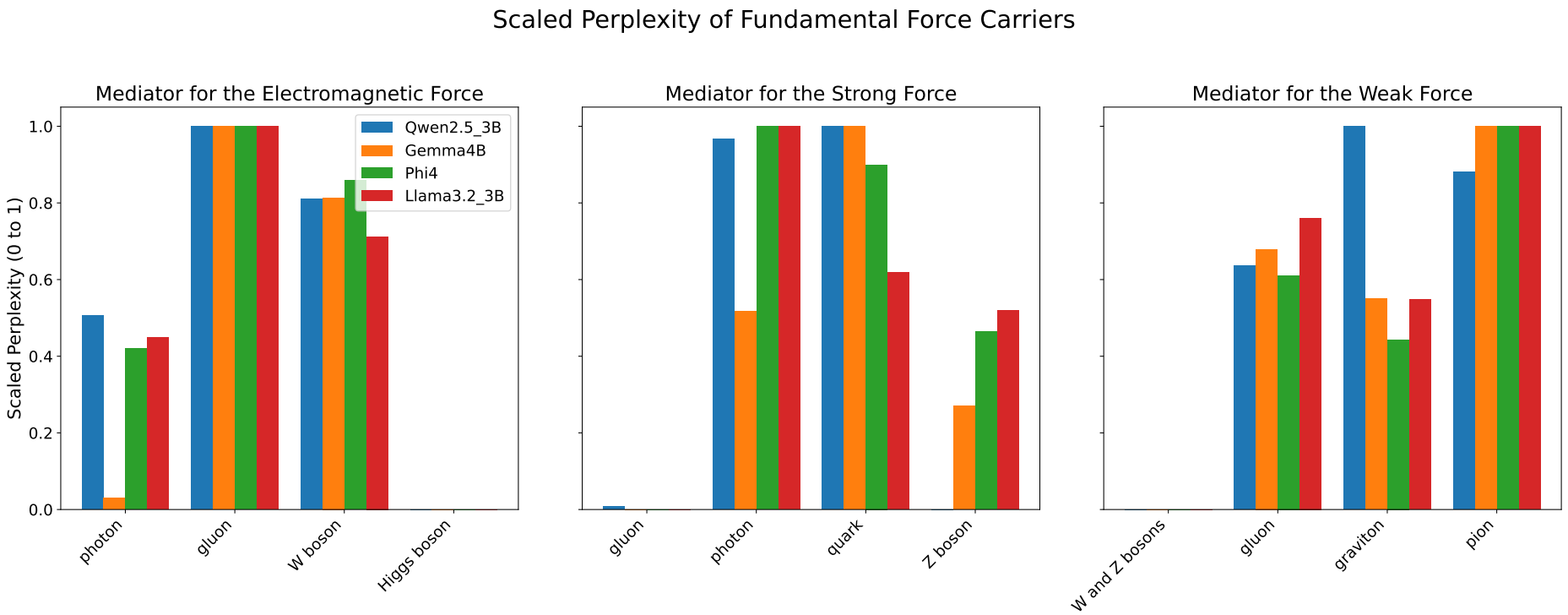
"Based on its total spin, a baryon is classified as {classification}."



$\longrightarrow$ Llamma3_2B and Gemma4B make small mistakes

Identifying the mediator of fundamental interactions

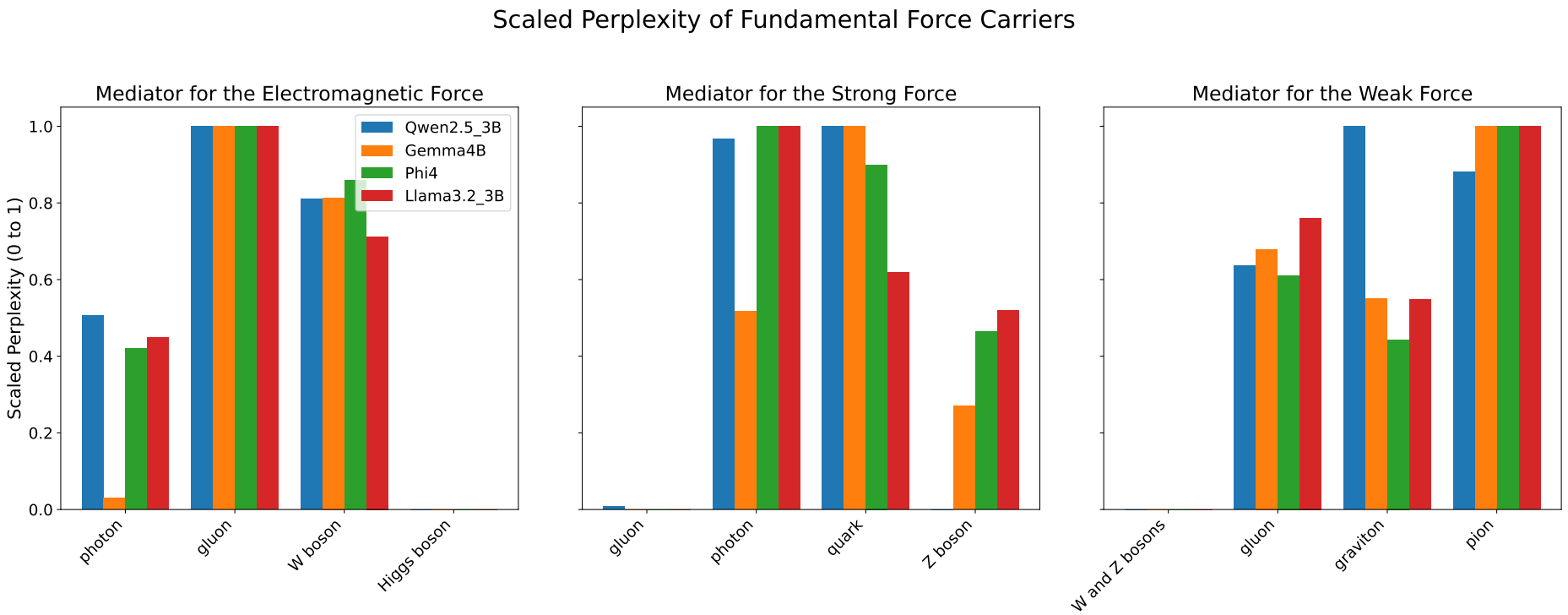
"The interaction of the {force} is mediated by {carrier}."



$\longrightarrow$ unexpected results for the e.m. force...

Identifying the mediator of fundamental interactions

"The interaction of the {force} is mediated by {carrier}."



$\longrightarrow$ unexpected results for the e.m. force...

Setting up a tool

- Online generators already exist and generating texts has become common practice.
- However, it is not without pitfalls: it is tempting to generate text and to forget to carefully check it, whereas they clearly do mistakes.
- Whether you write the text by yourself or use a generator, your text needs proof-checks.

Application: provide an AI companion to help proof-check scientific manuscripts:

1. Test and benchmark on small scale $\longrightarrow$ *done*
2. Same on large scale $\longrightarrow$ *ongoing*
3. Deploy a tool to analyze PDFs $\longrightarrow$ *to do*

https://tinyurl.com/EPSHEP/

Setting up a tool

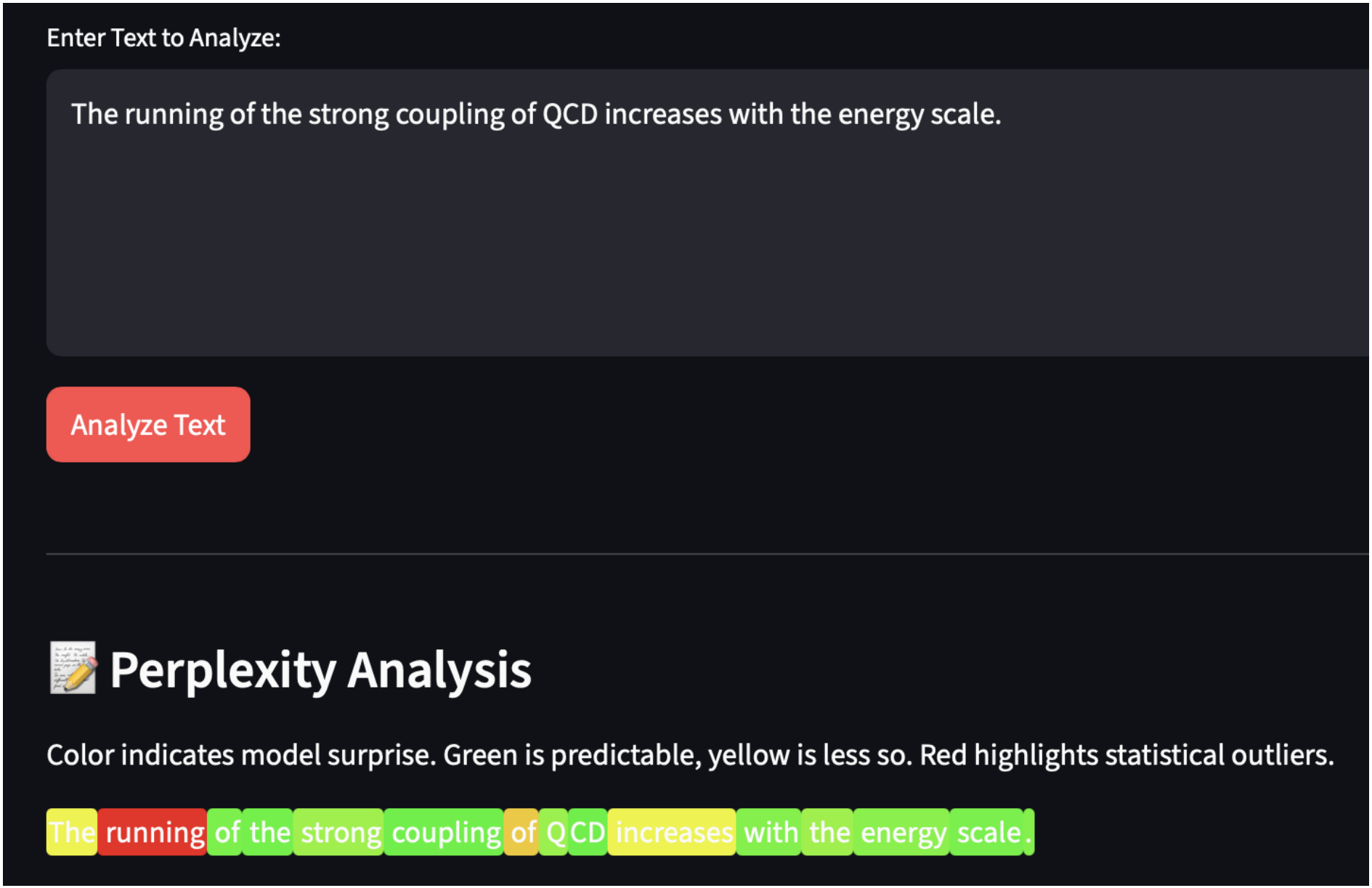
- Online generators already exist and generating texts has become common practice.
- However, it is not without pitfalls: it is tempting to generate text and to forget to carefully check it, whereas they clearly do mistakes.
- Whether you write the text by yourself or use a generator, your text needs proof-checks.

`https://tinyurl.com/EPSHEP/`

Application: provide an AI companion to help proof-check scientific manuscripts:

1. Test and benchmark on small scale $\longrightarrow$ *done*
2. Same on large scale $\longrightarrow$ *ongoing*
3. Deploy a tool to analyze PDFs $\longrightarrow$ *to do*

Online tool


Enter Text to Analyze:
The running of the strong coupling of QCD increases with the energy scale.
Analyze Text
Perplexity Analysis
Color indicates model surprise. Green is predictable, yellow is less so. Red highlights statistical outliers.
The running of the strong coupling of QCD increases with the energy scale.

Summary & Prospects

- We have investigated the knowledge of publicly available LLMs in high energy physics.
- Open-weight models perform reasonably well but still make certain mistakes.
- No model performs significantly better than others.
- However, LLMs are good enough to assist human beings at writing text, and we have set up an online prompt that you are welcome to test / challenge.

Thanks for your attention!

Summary & Prospects

- We have investigated the knowledge of publicly available LLMs in high energy physics.
- Open-weight models perform reasonably well but still make certain mistakes.
- No model performs significantly better than others.
- However, LLMs are good enough to assist human beings at writing text, and we have set up an online prompt that you are welcome to test / challenge.

Thanks for your attention!

Probing the model's knowledge of quark mass scales

Question: is the model's knowledge of a particle's mass influenced by the context?
Test: We test the prompt for a specific quark (e.g., the top quark) not just at its own energy scale, but across wildly different energy scales, from the MeV-range up to the TeV-range.

"The mass scale of the quark is {energy} GeV"

Hypothesis: The model's perceived mass (the point of lowest perplexity) will be context-dependent. The minimum perplexity will shift based on the energy scale being scanned, indicating its knowledge is not a fixed value but is biased by the prompt's context.

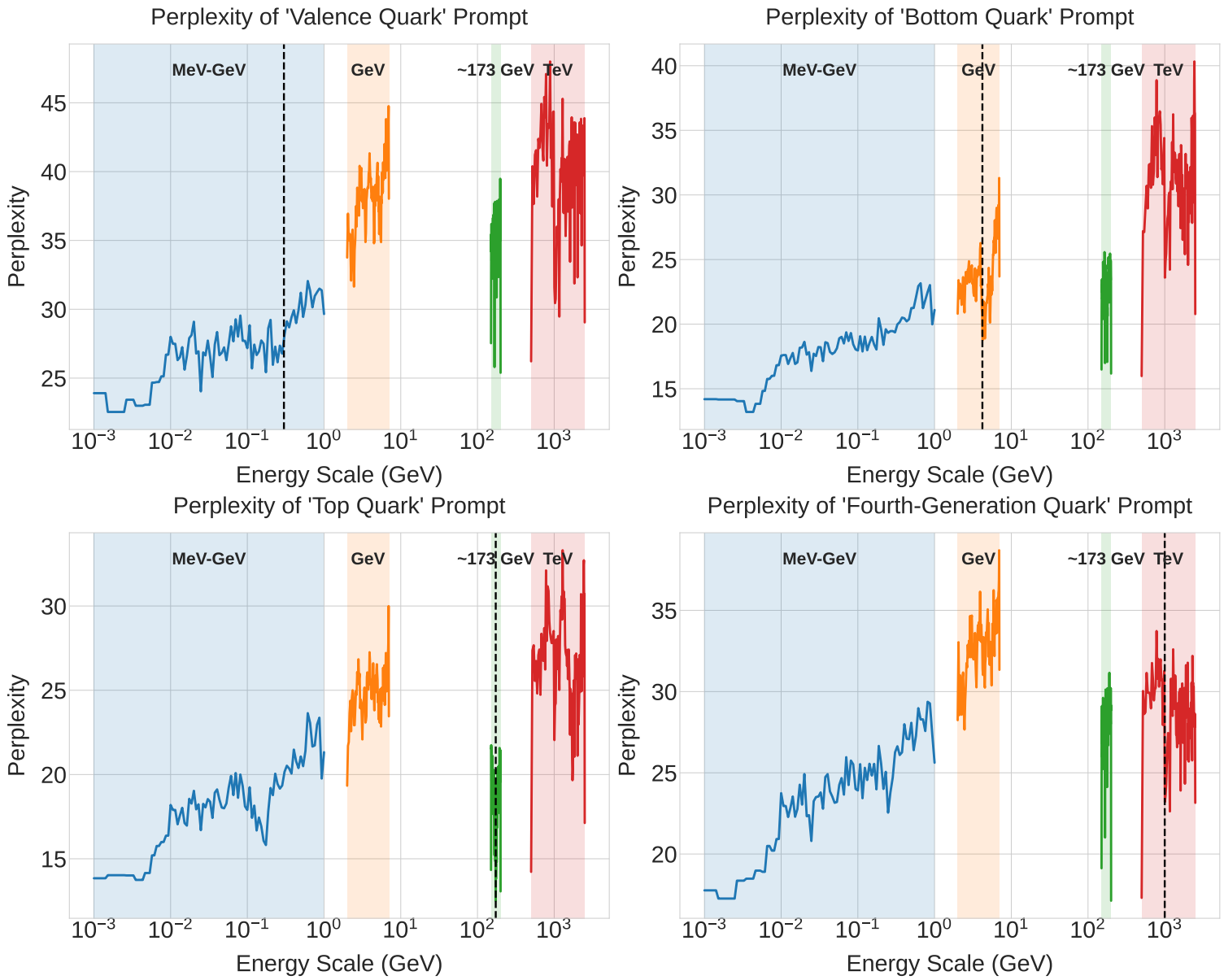
Probing the model's knowledge of quark mass scales

Question: is the model's knowledge of a particle's mass influenced by the context?
Test: We test the prompt for a specific quark (e.g., the top quark) not just at its own energy scale, but across wildly different energy scales, from the MeV-range up to the TeV-range.

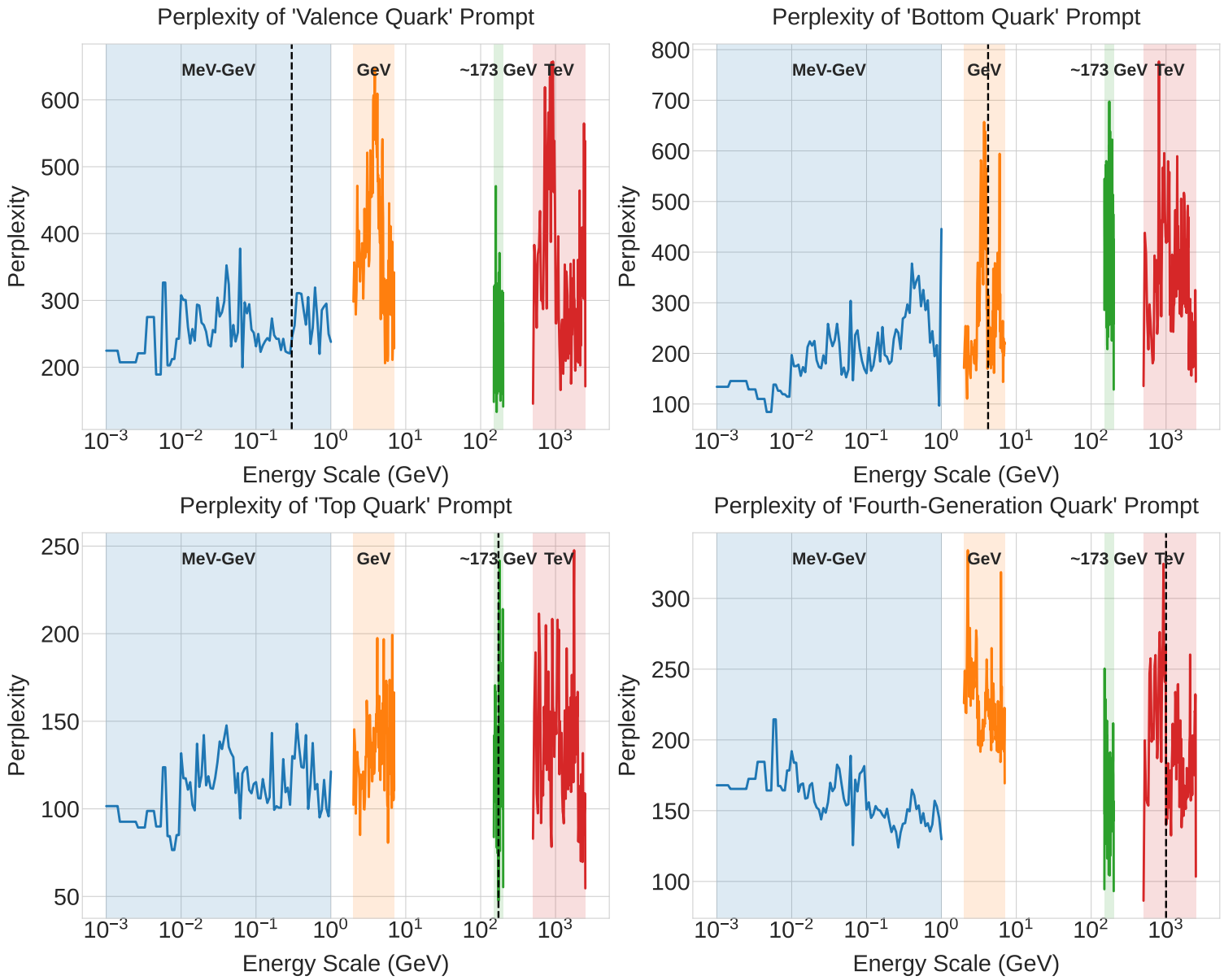
"The mass scale of the quark is {energy} GeV"

Hypothesis: The model's perceived mass (the point of lowest perplexity) will be context-dependent. The minimum perplexity will shift based on the energy scale being scanned, indicating its knowledge is not a fixed value but is biased by the prompt's context.

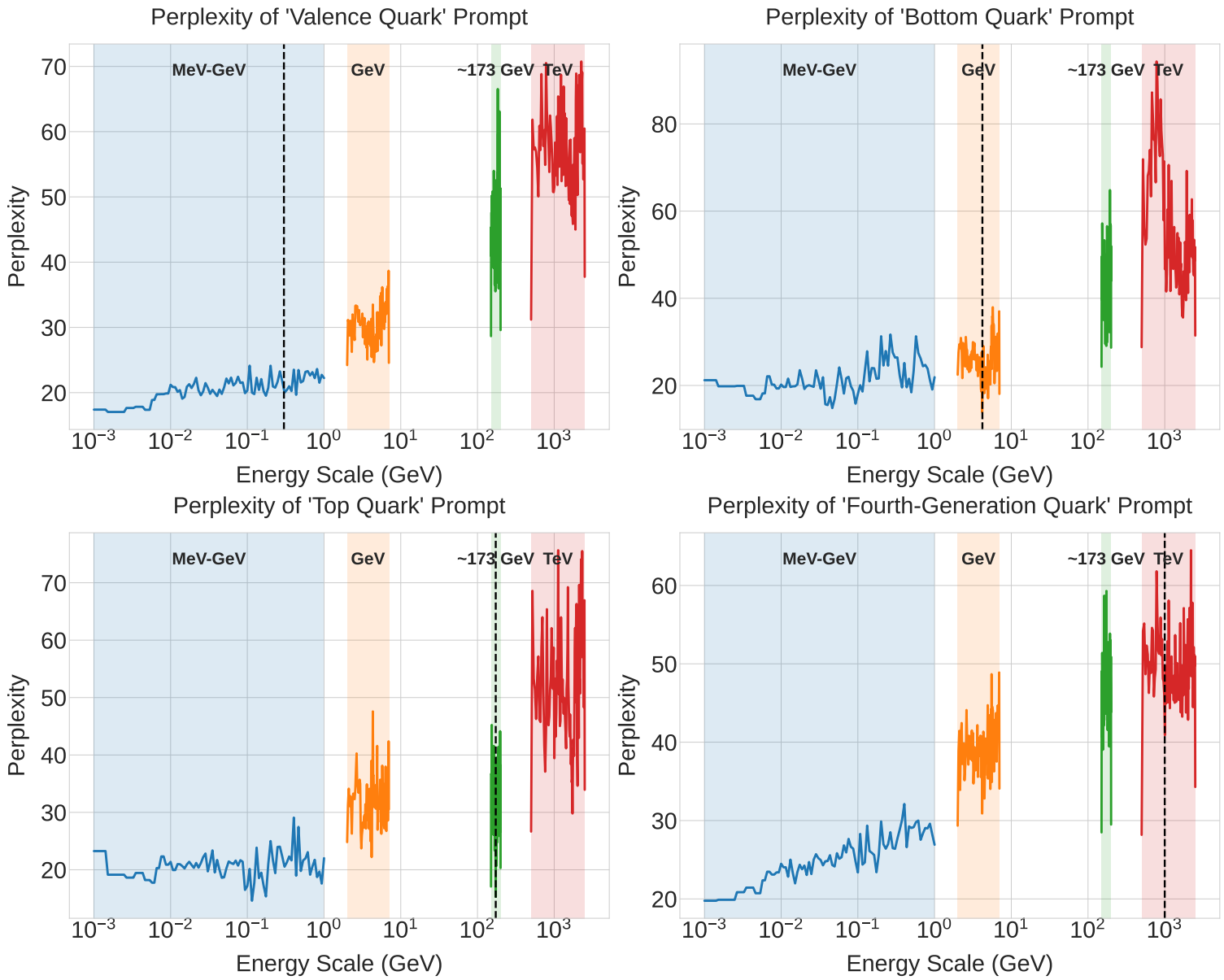
Qwen2.5-3B-Instruct LLM



Gemma4B-Instruct LLM



Phi4-Instruct LLM



Llama3.2-3B LLM