Hierarchical Prompting Taxonomy: A Universal Evaluation Framework for Large Language Models Aligned with Human Cognitive Principles

Devichand Budagam[1] , Ashutosh Kumar[2], Mahsa Khoshnoodi[3], Sankalp KJ[4], Vinija Jain[5], Aman Chadha[6]

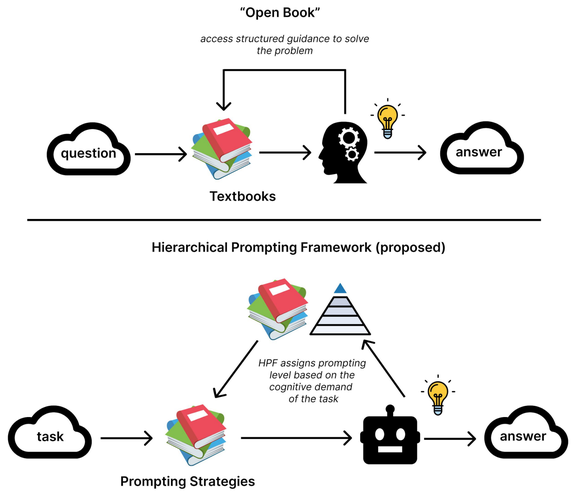
[1]Indian Institute of Technology Kharagpur, India [2]Rochester Institute of Technology, USA
[3]Researcher, Fatima Fellowship [4]AI Institute, University of South Carolina, USA
[5]Meta AI, USA, [6]Amazon GenAI USA

Association for the Advancement of Artificial Intelligence

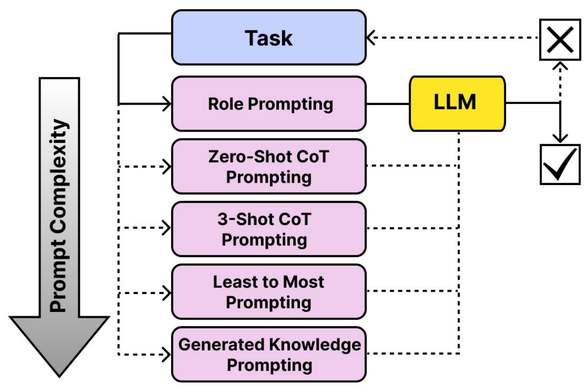
Scan here to check out our paper!

Introduction & Motivation

Modern evaluations of Large Language Models (LLMs) often rely on surface-level metrics that overlook the depth of reasoning these models can achieve. By drawing inspiration from human cognitive processes, our work introduces the **Hierarchical Prompting Taxonomy (HPT)** and **Hierarchical Prompting Framework (HPF)** to systematically decompose tasks—from basic recall to complex reasoning and knowledge integration. Using the **Hierarchical Prompting Index (HPI)** to quantify cognitive demands, our approach aligns model evaluation with human cognitive principles, enabling adaptive prompting strategies that enhance both performance and interpretability across a range of applications.



Analogical framework comparing the HPF with "Open Book" examination methodology. The diagram illustrates how HPF components (below) mirror traditional educational assessment elements (above), with parallel relationships between task complexity levels, resource utilization (prompts/textbooks), and performance metrics (HPI/student effort). This comparison demonstrates how LLM task complexity scales similarly to educational assessment complexity, from simple lookup tasks to complex synthesis problems

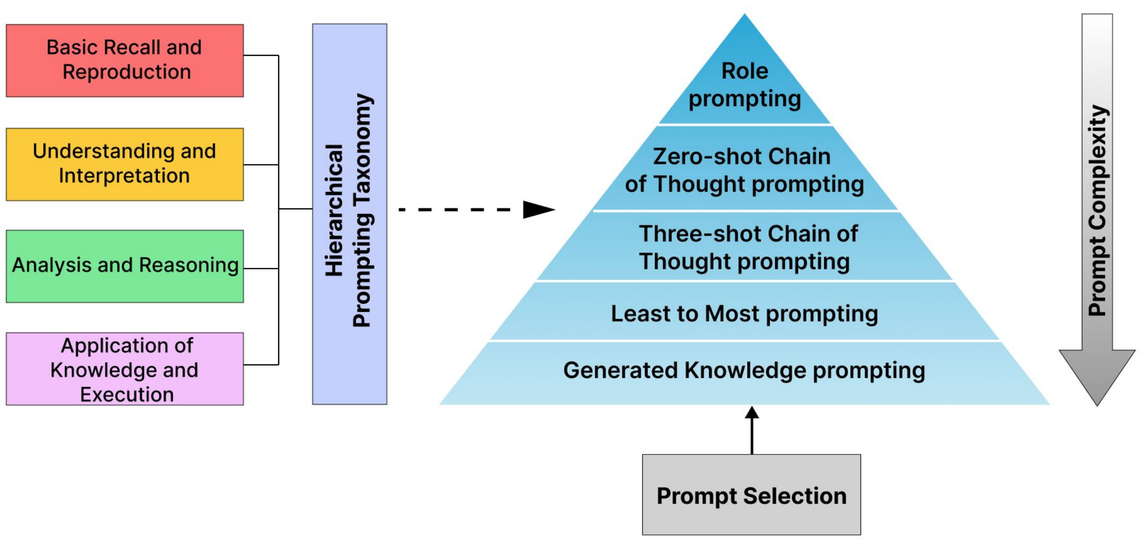


Hierarchical Prompting Framework includes five distinct prompting strategies, each designed for different levels of task complexity to ensure the appropriate prompt is selected for the given task. A ✓ indicates task completion, while a × signifies task incompletion.

Research Questions

1. How can HPF be used to match prompt complexity to task cognitive demands, optimizing the trade-off between computational efficiency and performance?
2. Under what conditions can strategic prompt selection enable smaller language models to achieve comparable performance to larger models, and how does this relate to task complexity as measured by Hierarchical Prompting Index (HPI)?
3. How can the measurements of task cognitive demands through HPI inform model selection and deployment decisions, beyond traditional accuracy metrics?

Hierarchical Prompting Taxonomy (HPT)



Hierarchical Prompting Taxonomy: A taxonomy designed to assess the complexity of prompting strategies based on the criteria: Basic Recall and Reproduction, Understanding and Interpretation, Analysis and Reasoning, and Application of Knowledge and Reasoning.

Hierarchical Prompting Framework (HPF)

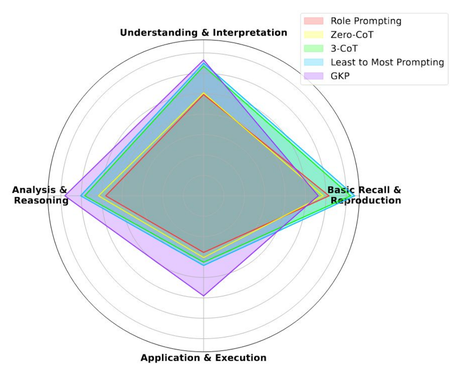
An operational system that implements HPT by sequentially applying prompting strategies of increasing complexity. HPF progressively challenges a language model to solve tasks, starting with minimal cognitive demands and advancing to multi-step reasoning as needed.

Hierarchical Prompting Index (HPI)

A metric that quantifies the cognitive effort required by a language model to solve a given task. The HPI is determined by the prompting level at which the model first produces a correct response, with lower values indicating easier tasks and higher values reflecting increased cognitive load **(1 being the lowest and 5 being the highest)**.

Algorithm 1: HPI Metric

```
HPI_List = [ ]
for sample i in the evaluation dataset do
    for level x in the HPF do
        if LLM resolves the task then
            HPI_List[i] = x
            break
        end if
    end for
    if LLM failed to resolve the task then
        HPI_List[i] = m + HPI_Dataset
    end if
end for
HPI = (1/n) Σ_{j=1}^{n} HPI_List[j]
```



Scoring distribution for each of the four rules of the HPT–*Basic Understanding & Reproduction, Understanding & Interpretation, Analysis and Reasoning, Application of Knowledge & Execution* for the prompting strategies in the HPF

Experimental Setup & Evaluation

Dataset	Evaluation Set Size	Representative Set Size	$HPI_{Dataset}$
MMLU	14500	725	3.03
GSM8k	1320	66	2.14
Humaneval	160	8	4.68
BoolQ	3270	162	1.71
CSQA	1221	60	2.52
IWSLT	890	45	1.92
SamSum	819	40	2.23

$HPI_{Dataset}$ scores across datasets evaluated by human annotators). The table lists the evaluation set size, representative set size, and $HPI_{Dataset}$ for various datasets. $HPI_{Dataset}$ scores provide a measure of task complexity relative to human annotators.

Proprietary Models:
GPT-4o, Claude 3.5 Sonnet

Open-Source Modes:
Gemma 7B, Mistral 7B, Llama-3 8B, Gemma-2 9B, Mistral-Nemo 12B

Task Categories:
Reasoning, Coding, Mathematics, Question-Answering, Summarization, Machine Translation

Results

DATASETS	MMLU		GSM8k		BoolQ		CSQA	
Models	HPI	Accuracy	HPI	Accuracy	HPI	Accuracy	HPI	Accuracy
GPT-4o	1.81	91.61	1.71	96.43	1.32	96.82	**1.65**	**92.54**
Claude 3.5 Sonnet	1.84	92.16	**1.35**	**97.72**	1.20	99.81	2.01	86.15
Mistral-Nemo 12B	2.45	89.75	3.01	86.80	1.75	99.87	2.06	90.17
Gemma-2 9B	2.34	87.28	2.17	91.28	1.30	98.28	1.94	88.86
Llama-3 8B	2.84	82.63	2.34	86.20	1.37	99.30	2.43	84.76
Gemma 7B	2.93	83.31	6.70	27.88	1.45	99.42	2.50	83.78
Mistral 7B	2.89	81.45	5.11	46.93	1.41	98.07	2.49	82.06

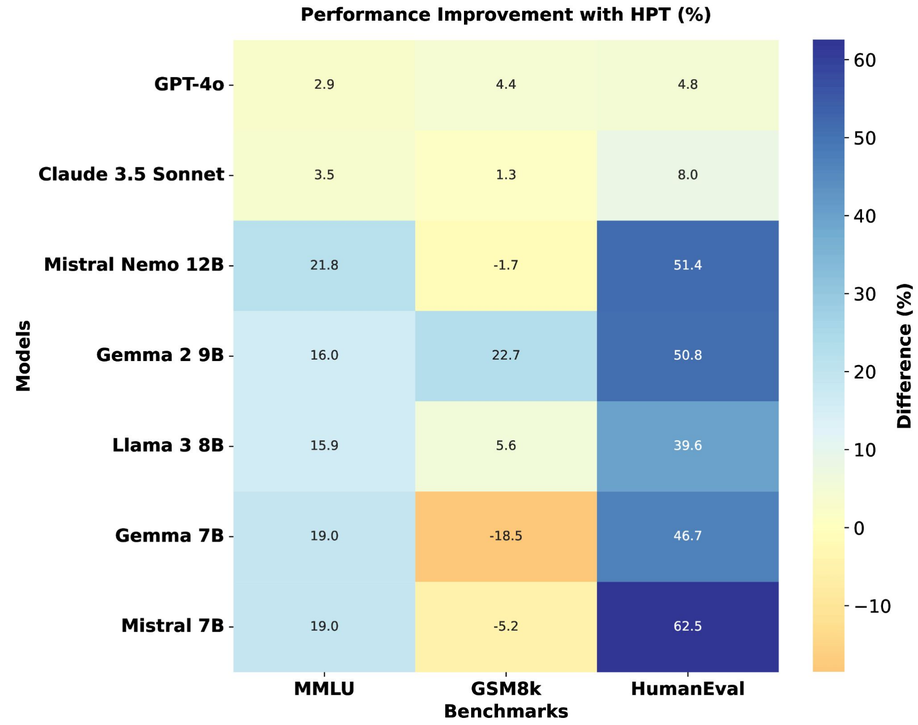
HPI (lower is better) and accuracy of LLMs across MMLU, GSM8K, BoolQ, and CSQA datasets. Blue indicates datasets where the LLM with the best HPI does not achieve the best performance. Green indicates the LLM with the best performance over the maximum number of datasets.

DATASETS	IWSLT				SamSum			
	HPI		BLEU		HPI		ROUGE-L	
Models	0.15	0.20	0.15	0.20	0.15	0.20	0.15	0.20
GPT-4o	**2.66**	**3.08**	**0.32**	**0.32**	**1.11**	**1.21**	**0.30**	**0.29**
Claude 3.5 Sonnet	4.63	4.87	0.20	0.20	1.25	1.60	0.23	0.23
Mistral-Nemo 12B	2.87	3.40	0.27	0.27	1.19	1.47	0.23	0.24
Gemma-2 9B	4.40	4.75	0.21	0.20	1.30	1.86	0.22	0.22
Llama-3 8B	3.40	3.92	0.24	0.23	1.30	1.72	0.22	0.22
Gemma 7B	5.39	5.84	0.08	0.09	3.31	5.03	0.11	0.10
Mistral 7B	3.52	4.14	0.22	0.22	1.26	1.68	0.21	0.22

HPI (lower is better), BLEU score for IWSLT, and ROUGE-L score for SamSum, of LLMs with thresholds.

DATASET	HumanEval	
Models	HPI	Pass@1
GPT-4o	2.25	0.95
Claude 3.5 Sonnet	1.04	1.00
Mistral-Nemo 12B	2.07	0.96
Gemma-2 9B	1.01	0.91
Llama-3 8B	1.03	1.00
Gemma 7B	3.71	0.79
Mistral 7B	1.10	0.93

HPI (lower is better) and Pass@1 of LLMs on the HumanEval dataset. Blue indicates datasets where the LLM with the best HPI does not achieve the best performance



Performance Comparison of HPT-based Evaluation vs. Standard Evaluation}: Performance improvements (in \%) when using HPT-based evaluation compared to standard evaluation across three benchmarks: MMLU, GSM8k, and HumanEval. Positive values indicate performance gains with HPT, while negative values indicate performance decreases. The baseline standard evaluation scores are sourced from Hugging Face leaderboard and official research reports.

Discussions

Cognitive Load-Based Prompt Selection: HPF improved performance with models showing upto 21.8% improvement on MMLU benchmarks

Model-Prompt Efficiency Trade-offs: Enabled smaller models to achieve better performance than larger models for specialized tasks

Cognitive Load as Model Selection Criterion: Models with similar accuracy scores often showed different HPI values, revealing cognitive effort requirements. While Claude 3.5 achieved highest MMLU accuracy, GPT-4o recorded the best HPI score, demonstrating that cognitive load provides valuable insights beyond traditional metrics.

Conclusions

HPT effectively assesses LLMs by focusing on cognitive task demands and using tailored prompting strategies, leading to improved performance across datasets. It provides insights into LLM problem-solving and suggests that dynamic prompting enhances evaluation methods. This approach aligns evaluation with human cognitive principles, paving the way for better benchmarks and in-context learning methods.

Acknowledgements

This material is based upon work partially supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.