



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Informatics

Metacognition and critical thinking: Instance-level demand
scales and annotation by LLMs

End of Degree Project

Bachelor's Degree in Data Science

AUTHOR: Jiang Chen, Kexin

Tutor: Hernández Orallo, José

Cotutor: Martínez Plumed, Fernando

ACADEMIC YEAR: 2023/2024

Resum

El camp de la intel·ligència artificial ha portat al desenvolupament de grans models de llenguatge avançats amb impressionants habilitats lingüístiques. No obstant això, encara no està clar fins a quin punt aquests models posseeixen habilitats metacognitives, les quals són fonamentals per a un raonament i aprenentatge avançats. Aquest projecte té com a objectiu avaluar les habilitats de metacognició i pensament crític en els grans models de llenguatge, amb un enfocament en la identificació de les escales i dimensions més efectives per a l'avaluació. Proposem un marc integral, que abasta tres dimensions clau: la necessitat de processos de pensament crític, la dificultat de calibrar el conegut i l'inconegut, i la dificultat en identificar informació rellevant. Aquest marc s'utilitza per anotar instàncies de preguntes en diversos benchmarks de BIG-Bench i HELM, destinats a mesurar habilitats cognitives avançades en els grans models de llenguatge. Les anotacions són generades pel model de llenguatge de última generació, GPT-4. Aquestes anotacions s'utilitzen després com a predictors per a construir models de rendiment per a diversos grans models de llenguatge en aquests benchmarks, amb l'objectiu final de determinar fins a quin punt aquests benchmarks realment mesuren les capacitats metacognitives. Els nostres resultats mostren que, mentre que molts models manquen de capacitats metacognitives, els models més grans mostren algunes indicacions de tals habilitats. A més, l'ús d'una escala multidimensional per a les demandes metacognitives millora la predictibilitat del rendiment en comparació amb una escala integrada única. En proporcionar una eina d'avaluació per a la metacognició en els grans models de llenguatge, aquest projecte ofereix informació sobre l'efectivitat dels benchmarks en l'avaluació de les habilitats metacognitives. Els resultats destaquen la importància d'un disseny meticulós dels benchmarks i el potencial de les escales multidimensionals per capturar la naturalesa complexa de la metacognició.

Paraules clau: Avaluació d'IA, GPT4, grans models de llenguatge, metacognició, pensament crític, predictibilitat.

Resumen

El campo de la inteligencia artificial ha llevado al desarrollo grandes modelos de lenguaje avanzados con impresionantes habilidades lingüísticas. Sin embargo, aún no está claro hasta qué punto estos modelos poseen habilidades metacognitivas, las cuales son fundamentales para un razonamiento y aprendizaje avanzados. Este proyecto tiene como objetivo evaluar las habilidades de metacognición y pensamiento crítico en los grandes modelos de lenguaje, con un enfoque en la identificación de las escalas y dimensiones más efectivas para la evaluación. Proponemos un marco integral, que abarca tres dimensiones clave: la necesidad de procesos de pensamiento crítico, la dificultad de calibrar lo conocido y lo desconocido, y la dificultad en identificar información relevante. Este marco se utiliza para anotar instancias de preguntas en varios benchmarks de BIG-Bench y HELM, destinados a medir habilidades cognitivas avanzadas en los grandes modelos de lenguaje. Las anotaciones son generadas por el modelo de lenguaje de última generación, GPT-4. Estas anotaciones se utilizan luego como predictores para construir modelos de rendimiento para varios grandes modelos de lenguaje en estos benchmarks, con el objetivo final de determinar hasta qué punto estos benchmarks realmente miden las capacidades metacognitivas. Nuestros resultados muestran que, mientras que muchos modelos carecen de capacidades metacognitivas, los modelos más grandes muestran algunas indicaciones de tales habilidades. Además, el uso de una escala multidimensional para las demandas metacognitivas mejora la predictibilidad del rendimiento en comparación con una escala integrada única. Al proporcionar una herramienta de evaluación para la metacognición en los grandes modelos de lenguaje, este proyecto ofrece información sobre la efectividad de los benchmarks en la evaluación de las habilidades metacognitivas. Los resultados destacan la importancia de un diseño meticuloso de los benchmarks y el potencial de las escalas multidimensionales para capturar la naturaleza compleja de la metacognición.

Palabras clave: Evaluación de IA, GPT4, grandes modelos de lenguaje, metacognición, pensamiento crítico, predictibilidad.

Abstract

The rapidly growing field of artificial intelligence has led to the development of advanced Large Language Models (LLMs) with impressive language skills. However, it is still unclear the extent to which these models possess metacognitive abilities, which are critical for advanced reasoning and learning. This study aims to evaluate metacognition and critical thinking abilities in LLMs, with a focus on identifying the most effective scales and dimensions for assessment. We propose a comprehensive framework, encompassing three key dimensions: the need for critical thinking processes, the difficulty of calibrating knowns and unknowns, and the difficulty in identifying relevant information. This framework is used to annotate question instances across several benchmarks from BIG-Bench and HELM, aimed at measuring advanced cognitive skills in LLMs. The annotations are generated by the state-of-the-art LLM, GPT-4. These annotations are then used as predictors to build performance models for various LLMs on these benchmarks, with the ultimate goal of determining the extent to which the benchmarks truly measure metacognitive capabilities. Our findings reveal that while many models lack metacognitive capabilities, larger models exhibit some indications of such abilities. Furthermore, the use of a multi-dimensional scale for metacognitive demands improves the predictability compared to a single integrated scale. By providing an evaluation tool for metacognition in LLMs, this study provides insights into the effectiveness of benchmarks in assessing metacognitive abilities. The finding highlights the importance of careful benchmark design and the potential of multi-dimensional scales in capturing the complex nature of metacognition.

Key words: AI evaluation, GPT-4, Large Language Models, metacognition, critical thinking, predictability

Contents

Contents	vii
List of Figures	ix
List of Tables	ix
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	2
1.3 Structure	2
2 Background	5
2.1 Metacognition and Critical Thinking	5
2.2 Large Language Models (LLMs)	6
2.2.1 Automatic Annotation using LLMs	7
2.3 Assessors	7
3 Data	9
3.1 BIG-bench	9
3.2 HELM	11
3.3 Scenarios of Assessment	13
4 Scales and Rubrics	15
4.1 Demands definition	15
4.2 Rubrics	17
4.2.1 Three Metacognition Scale	17
4.2.2 One Metacognition Scale	20
4.2.3 General Difficulty	22
4.3 Human Labelling	24
5 Experimental setting	29
5.1 Data processing	29
5.2 Large Language Models (LLMs)	30
5.3 Assessors	32
5.3.1 Model Configuration	32
5.4 Evaluation Scenarios	34
5.5 Evaluation Metrics	35
5.5.1 Area Under the Receiver Operating Characteristic Curve (AUC)	35
5.5.2 Brier Score	35
6 Results	37
6.1 Overview	37
6.2 Metacognition-loaded datasets	38
6.2.1 Metacognition-demands scale	38
6.2.2 General difficulty scale	42
6.3 Contrast datasets	43
6.3.1 Metacognition-demands scale	44
6.3.2 General difficulty scale	46
7 Conclusions	49

7.1	Limitations and Future Work	50
7.2	Legacy	51
8	Acknowledgements	53
	Bibliography	55

Appendices		
A	Three Metacognition Scales	59
A.1	Metacognition-loaded datasets	59
A.1.1	GPT Family	59
A.2	Contrast datasets	60
A.2.1	Original HELM models	60
A.2.2	GPT Family	61
B	Sustainable Development Goals	63

List of Figures

5.1	Distribution of instances across the different levels of each scale: Metacognition-demands scale and general difficulty scale, for each of the benchmarks from both metacognition-loaded datasets and contrast datasets.	30
5.2	Distribution of instances across the different levels of each scale: Metacognition-demands scale and general difficulty scale, for both the metacognition-loaded datasets and contrast datasets, with all benchmarks aggregated. . .	31
6.1	Success rate over the level of demand on the metacognition-loaded datasets using each scale: metacognition-demands scale and general difficulty scale, when using GPT3-5 and GPT-4 to generate instance outputs. The black line represents the aggregated success rate across all benchmarks, and the the colored lines represent individual benchmarks with a confidence interval of 95%. The dotted line for the metacognition-demands scale represents an extrapolation for the level 5 metacognition demand, calculated using a linear regression of the data points from levels 1 to 4.	39
6.2	Success rate over the level of demand on the contrast datasets using each scale: metacognition-demands scale and general difficulty scale, when using GPT3-5 and GPT-4 to generate instance outputs. The black line represents the aggregated success rate across all benchmarks, and the the colored lines represent individual benchmarks.	40
6.3	Success rate over metacognition demand on the contrast datasets using each scale: metacognition-demands scale and general difficulty scale, when using the instance outputs of the model Microsoft TNLGv2_530B. The black line represents the aggregated success rate across all benchmarks, and the the colored lines represent individual benchmarks	45

List of Tables

3.1	Number of instances for each metacognition-loaded benchmark from the BIG-Bench repository.	10
3.2	Number of instances for each non-metacognition-loaded benchmark from the HELM repository.	13
4.1	Spearman correlation matrix between the original demands from the Three metacognition scale and the integrated metacognition scale.	22
4.2	Spearman correlation matrix between level of metacognition demand using the integrated metacognition scale and the performance of the BIG-G language models.	23

4.3	Global spearman correlation between human raters.	25
4.4	Spearman correlation between human raters for D1 (Need for critical thinking processes).	25
4.5	Spearman correlation between human raters for D2 (Difficulty in calibrating known and unknowns).	26
4.6	Spearman correlation between human raters for D3 (Difficulty in identifying relevant information).	26
4.7	Mean correlation for each of the three demands and combined.	26
4.8	Spearman correlation between GPT-4 and human raters across all demands and for each demand separately.	26
4.9	Spearman Correlation between Persona vs GPT-4.	27
5.1	Extracted BIG-bench models. Each model size value is a different model from the model family.	32
5.2	Extracted HELM models with their number of parameters.	33
6.1	Summary of assessor performance on the metacognition-loaded datasets evaluated using the metacognition-demands scale and the general difficulty scale, and using the instance outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.	38
6.2	Summary of assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale and the general difficulty scale, and using the instance outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.	38
6.3	Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs combination of all BIG-G models instance outputs, with results aggregated across all benchmarks.	40
6.4	Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs from the combination of all BIG-G models, with results disaggregated by individual benchmarks.	41
6.5	Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs of the model BIG-G 128b, with results aggregated across all benchmarks.	41
6.6	Assessor performance on the metacognition-loaded datasets evaluated using the three metacognition scale and one metacognition scale, and using the outputs of the model BIG-G 128b, with results disaggregated by individual benchmarks.	41
6.7	Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.	41
6.8	Assessor performance on the metacognition-loaded datasets evaluated using the metacognition-demands scale, with the outputs of the combination of the BIG-G models and the model BIG-G 128b, with results disaggregated by individual benchmarks.	42
6.9	Assessor performance on the metacognition-loaded datasets evaluated using the generic difficulty scale, with the outputs of the combination of the BIG-G models and the model BIG-G 128b, with results aggregated across all benchmarks.	42

6.10	Assessor performance on the metacognition-loaded datasets evaluated using the generic difficulty scale, with the outputs of the combination of BIG-G models and the model BIG-G 128b, with results disaggregated by individual benchmarks.	43
6.11	Assessor performance on the metacognition-loaded datasets evaluated using the general difficulty scale, and using the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.	43
6.12	Assessor performance on the metacognition-loaded datasets evaluated using the general difficulty scale, and using the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks. . . .	43
6.13	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results aggregated across all benchmarks.	44
6.14	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results disaggregated by individual benchmarks.	45
6.15	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.	46
6.16	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.	46
6.17	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results aggregated across all benchmarks.	47
6.18	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results disaggregated by individual benchmarks.	47
6.19	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.	48
6.20	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.	48
A.1	Assessor performance on the metacognition-loaded datasets evaluated using the three metacognitions scale, with the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.	59
A.2	Assessor performance on the metacognition datasets evaluated using the three metacognition scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.	60
A.3	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the three metacognition scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results aggregated across all benchmarks.	60

A.4	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the three metacognition scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results disaggregated by individual benchmarks.	61
A.5	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.	61
A.6	Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the three metacognition scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.	62

CHAPTER 1

Introduction

In recent years, there has been significant advancements in the field of Artificial Intelligence (AI) and Large Language Models (LLMs), which have demonstrated profound capabilities in processing and generating human-like text. These models can imitate conversational styles, understand context and generate coherent responses, making them incredibly valuable in applications ranging from automating customer service to supporting complex decision-making processes. Despite these advances, there remains a significant challenge in the field of AI, which is the ability to perform self-efficacy tasks similar to human metacognition, including self-evaluation and critical thinking.

Metacognition is a higher-order cognitive ability that enables individuals to monitor, control, and adapt their cognitive processes to the demands of a task. In humans, this ability is critical for effective learning and problem solving. Applying these processes to AI means developing systems that are not only self-aware, but also capable of assessing their performance and adapting their strategies in complex situations. The gap in implementing these self-regulatory functions in AI systems highlights a significant limitation that limits their applicability in scenarios that require autonomous adaptability and critical decision-making.

This work aims to fill this gap by proposing a systematic framework for evaluating and measuring the metacognitive demands of various AI tasks. The main contribution is the development of a predictive framework that uses a newly created metacognitive rubric to estimate how well AI systems can handle unseen tasks that require these higher-order thinking skills. This is achieved by analysing different instances of tasks designed to test metacognitive skills, and then quantifying the complexity and demands of these tasks.

1.1 Motivation

This project is conducted in collaboration with the OECD AI Skills Reasoning Group¹, driven by the need to systematically quantify the demands of metacognition and critical thinking in various question instances. The primary motivation behind this research is to develop a robust framework that can predict the performance of AI systems on new, unseen instances by accurately measuring these metacognitive demands.

The main focus of this project is the development and application of a comprehensive rubric designed to assess the metacognitive demands of tasks. This rubric will serve as a critical tool in evaluating whether existing benchmarks for metacognition truly capture the essence of this complex skill. By achieving this, we aim to ensure that AI systems are

¹<https://oecd.ai/en/work-innovation-productivity-skills>

not only advanced in their technical capabilities but also exhibit a level of thinking and reasoning comparable to human beings.

By systematically quantifying metacognitive demands, we can improve our understanding of how AI systems process and respond to complex tasks. This will allow us to develop more sophisticated AI models that are better equipped to handle real-world challenges. Ultimately, the goal of this project is to bridge the gap between human cognitive abilities and AI performance, fostering the development of AI systems that can think, reason, and adapt in ways that are more aligned with human intelligence.

1.2 Objectives ---

The primary objectives of this project are:

1. **Benchmark evaluation:** To evaluate existing metacognition benchmarks by analysing the performance of various language models. This involves determining whether these benchmarks are valid indicators of metacognitive abilities or if they mix together metacognitive demands with general task difficulty.
2. **Rubric development:** To develop and validate a comprehensive rubric for measuring metacognitive demands. This rubric should accurately reflect what the intended constructs and be capable of predicting language model performance. The goal is to ensure that the rubric can distinguish between tasks that require metacognitive skills and those that do not.
3. **Model comparison:** To compare the performance of different language models on metacognitive tasks. This includes analysing the evolution of metacognitive capabilities across different generations and architectures of language models, thereby identifying trends and improvements in AI metacognition.
4. **Scenario-based assessment:** To systematically assess the predictive power of metacognition demands and general difficulty on language model performance across different scenarios. This involves evaluating both metacognition-loaded and non-metacognition-loaded datasets using distinct scales.

1.3 Structure ---

The report is structured as follows:

Chapter 2 provides an overview of metacognition and critical thinking, as well as the role of Large Language Models (LLMs) in automatic annotation.

Chapter 3 details the benchmarks and datasets used in the study. It includes a description of the benchmark repositories, and explains the selection criteria for the datasets that align with our research objectives.

Chapter 4 defines the metacognitive demands and the development of the rubric. It discusses the different scales created, their prompt designs, and the validation process to ensure they accurately measure the intended constructs.

Chapter 5 outlines the experimental setup, including data processing, the selection of Large Language Models, and the configuration of assessors. It also describes the different scenarios investigated using the developed scales and benchmarks.

Chapter 6 presents the results of the experiments. It analyses the performance of various language models across different scenarios, comparing the effectiveness of the different scales in predicting model performance.

Chapter 7 summarizes the key findings of the research, discusses their implications, and suggests potential directions for future work.

CHAPTER 2

Background

2.1 Metacognition and Critical Thinking

Metacognition and critical thinking are a group of sophisticated capabilities that refer to "thinking about thinking" [1] [2]. In the field of artificial intelligence, these cognitive skills are increasingly recognised as essential for the development of robust AI systems and highlights the importance of metacognition in enabling AI systems to monitor their own learning processes, identify areas where they lack knowledge, and actively seek out additional information[3]. Miller (2019) [4] argues that critical thinking capabilities are necessary for AI systems to reason under uncertainty, evaluate the credibility of information, and generate justifications for their decisions.

Recent research shows how AI is progressing dramatically but focusing on narrow applications, therefore proposing to study human cognitive capabilities to enhance AI systems with skills like generalisability and ethical reasoning [5] based on D. Kahneman's theory of thinking fast and slow [6]. Similarly in another recent study, the authors propose a metacognitive approach as a safety measure, enabling AI systems to self-diagnose potential failures and adapt to new situations [7]. The rise of generative AI brings the question of metacognition to the forefront. By enabling these systems to understand and monitor their own actions, we can significantly improve their performance. The incorporation of a metacognition module in the generative agents can allow them self-reflect and enhance goal-directed behaviour[8], and can even go beyond with potential applications in psychology, education and interactive media[8].

However a lack of metacognitive abilities in AI can lead to challenges in various areas including ethical and societal impact, AI governance, and even bias and discrimination[9]. By incorporating metacognition and critical thinking into AI systems, researchers are moving towards the development of more flexible, adaptable, and ultimately trustworthy intelligent agents [10].

An important challenge has been highlighted discussing the understanding the degree of metacognition demand imposed on users [11]. However, a more significant challenge lies in estimating the level of metacognition and critical thinking these systems possess, especially when presented with new, unseen instances. While advancements have been made in areas like self-monitoring and justification generation, accurately gauging an AI's ability to adapt its thinking process or assess the validity of unfamiliar information remains an open question[12].

Evaluating metacognition in AI presents particular difficulties in terms of generalisability. Current metrics used to assess these capabilities may not translate well to unseen scenarios, as highlighted by Cox & Raja [13]. Furthermore, the potential for "super-

intelligence" to emerge, where AI surpasses human comprehension in its reasoning and decision-making, raises concerns about our ability to definitively assess its metacognitive abilities [14].

Metacognition is also associated with "calibration", the ability to determine how sure an individual is about their own knowledge. This is also common in machine learning, with metrics of confidence estimation quality based on calibration. A model is calibrated if for all the predictions or answers it says it is correct X% of the times, then it is actually correct X% of the times [14]. For instance, if we take all the questions the model says something is true with 70% confidence (or expresses a probability of 70% that it is true), then, if they are true 70% of the time, the model is perfectly calibrated for those questions [15]. Calibration requires binning multiple questions or calculating metrics such as the Brier score decomposition, apart from asking confidence to the respondent or access to the logprobs. Instead, we will explore more direct and general ways of evaluating metacognition and critical thinking.

2.2 Large Language Models (LLMs)

Large Language Models (LLMs) have transformed the natural language processing (NLP) field by generating cohesive and contextually appropriate text from a provided prompt [21]. These models have undergone training with large datasets, allowing them to forecast the likelihood of a word sequence and produce text that resembles human language patterns closely [22]. LLMs are built on the transformer model, using self-attention to analyze word sequences simultaneously and assess the importance of each word compared to the rest within a sentence [23]. LLMs are able to grasp long-distance relationships and comprehend the surrounding context in the text with greater efficiency [24]).

LLMs such as GPT (Generative Pre-trained Transformer) have demonstrated outstanding adaptability across a range of NLP tasks, including text generation, translation, summarizing, and question-answering [25]. Their ability to perform zero-shot or few-shot learning, where the model can understand and execute tasks with little to no task-specific training, has been particularly transformative [26]. For example, GPT-4, one of the most advanced LLMs, has shown remarkable proficiency in understanding and generating human-like text across various domains, achieving state-of-the-art performance in many benchmarks without task-specific fine-tuning [21].

The training of LLMs involves unsupervised learning on diverse datasets, which allows them to develop a broad understanding of language and general knowledge. This extensive training process, coupled with the transformer architecture, equips LLMs with the capability to generalize well to new tasks and domains [27]. However, the deployment of these models also raises concerns regarding computational cost, ethical implications, and potential biases embedded in the training data [28] [29].

In addition to their practical applications, LLMs have also contributed significantly to theoretical advancements in NLP. Research has shown that the self-attention mechanism in transformers not only improves performance but also provides insights into the inner workings of language models, such as how they represent syntactic and semantic information [30]. This has opened up new avenues for exploring the interpretability and explainability of LLMs, which are crucial for their responsible and transparent use in real-world applications [31].

2.2.1. Automatic Annotation using LLMs

Numerous machine learning models depend heavily on extensive datasets with labeled information to perform effectively. However, the task of labeling such extensive data can be both time-consuming and financially and computationally burdensome. This is particularly true for text annotation, which involves navigating the intricate nuances of natural language, making it a notably laborious and complex process.

LLMs have become increasingly proficient and are now commonly utilized as automatic annotators [45, 46, 47], a process where a model identifies and labels data within a text. For example, one can pose a prompt (or query) to an LLM, such as: "Does the following sentence contain a phrasal verb: [...]?" By tailoring this prompt for each specific sentence, it is possible to swiftly generate an annotated dataset. This dataset indicates whether each sentence includes a phrasal verb, serving as a foundational resource for subsequent analytical or processing tasks.

It is not surprising that one of the notable applications of GPT-4 is its use in automatic annotation. This efficiency and performance is critical as the quality and accuracy of automatic annotation directly affect the performance of downstream machine learning models that rely on this annotated data for training [32]. GPT-4's sophisticated capabilities make it particularly well-suited for this task due to its ability to understand context at a granular level and handle a wide range of tasks effectively [21]. Integrating GPT-4 into the automatic annotation process can significantly reduce the time and resources required for data labeling while simultaneously enhancing the quality of the resulting datasets [25].

The architecture of GPT-4 allows it to effectively parse and comprehend complex language structures, making it proficient at identifying subtle nuances in text that are crucial for accurate annotation [22]. This proficiency is further enhanced by its training on extensive and diverse datasets, which provide GPT-4 with a comprehensive understanding of various contexts and domains [27]. Moreover, GPT-4's ability to perform few-shot learning enables it to adapt to new annotation tasks with minimal additional training, enhancing its practicality in automatic annotation scenarios [26].

In practical applications, deploying GPT-4 for automatic annotation can streamline workflows in data-intensive fields such as biomedical research, legal document analysis, and content moderation. By automating the annotation process, organizations can allocate human resources to more complex tasks requiring deeper analysis and judgment, thereby optimizing overall productivity and efficiency [32]. Additionally, the high-quality annotations produced by GPT-4 can lead to the development of more accurate and robust machine learning models, as the initial training data is of superior quality [29].

2.3 Assessors

An assessor is a conditional probability estimator designed to predict the probability of success for AI systems at the instance granularity level. The concept of assessors as a means to predict scores using machine learning techniques was initially introduced in a general context by [35]. The approach was further explored specifically for LLMs in [48] and [49, Sec. 5], yielding encouraging outcomes.

In a nutshell, assessor models allow us to anticipate the level of success for a particular instance without running it through the AI system or language model at all. The primary function of an assessor is to analyze various features of a given task and the

corresponding model performance to predict the model’s effectiveness on new instances. By doing so, assessors can provide valuable insights into the strengths and weaknesses of different models, guiding further improvements and refinements.

In this context, we use assessors to estimate the likelihood of success for LLMs on metacognitive tasks. Metacognitive tasks often involve self-evaluation, awareness of knowledge gaps, and the ability to distinguish between essential and non-essential information. These tasks are inherently complex and require sophisticated evaluation mechanisms to accurately measure model performance [33]. To build effective assessors, we employed machine learning algorithms that can learn from instance-level data. This granular approach allows the assessors to capture detailed patterns and relationships within the data, leading to more accurate predictions [34]. The specific machine learning techniques and configurations used for training the assessors are discussed in detail in the Experimental Setting chapter.

CHAPTER 3

Data

In this chapter, we discuss the data collected as well the methodology used to conduct our analysis. Instance-level data is crucial for AI evaluation as it provides a detailed and granular view of a large language model’s performance on specific inputs. This allows for a more thorough and accurate assessment of the model’s capabilities, offering insights into how the model performs on individual examples rather than just overall statistics. Two primary task repositories, BIG-bench and HELM, serve as the foundation for evaluating the capabilities of large language models. Both repositories provide instance-level data for numerous language models, which is essential for our study.

3.1 BIG-bench

The Beyond the Imitation Game benchmark (BIG-bench) [36] is a collaborative benchmark designed to evaluate the capabilities of large language models (LLMs) across a wide range of tasks. The primary objective of BIG-bench is to probe the limits of LLMs and to extrapolate their potential future capabilities. This benchmark includes tasks from diverse domains such as common sense reasoning, algebra, causal reasoning, metacognition, and critical thinking, among others.

For the purpose of this project, we have identified and selected tasks within the BIG-bench repository that are particularly relevant to our focus on metacognition. Specifically, we concentrated on tasks labeled under "self-evaluation," "sufficient information," and "self-awareness." These labels indicate tasks that require the model to engage in higher-order thinking processes, such as evaluating its own knowledge and reasoning capabilities.

From the extensive list of tasks available in BIG-bench, we have chosen three tasks that align closely with our research objectives, as well as their feasibility for implementation in our study. These tasks are:

1. **Evaluating Information Essentiality (EIE):** This task assesses a language model’s ability to identify which statements are essential for answering a given question. It measures the model’s capacity to discern critical information from unnecessary details in the context of the question posed, which is a key aspect of metacognitive processing.
2. **Known Unknowns (KU):** This task evaluates a model’s ability to recognize when it does not know the answer to a question. The model must correctly identify that the answer is unknown, which involves a metacognitive understanding of its own

knowledge limitations. This task is crucial for assessing the model’s self-awareness and its ability to avoid overconfident or incorrect responses.

3. **VitaminC Fact Verification (VFFV):** This task measures a model’s ability to verify the truthfulness of a given claim based on related external evidence. The model must infer the veracity of the claim by cross-referencing it with provided evidence, demonstrating critical thinking and fact-checking abilities.

Table 3.1 shows the number of instances for each task.

Benchmark	Number of Instances
Evaluating Information Essentiality (EIE)	68
Known Unknowns (KU)	46
VitaminC Fact Verification (VFFV)	66

Table 3.1: Number of instances for each metacognition-loaded benchmark from the BIG-Bench repository.

The number of instances for each selected task is relatively small, with a total of 180 instances across the three tasks. This limited sample size is due to the specific focus on metacognitive tasks within the BIG-bench repository, which inherently narrows the pool of available data.

While a larger sample size would provide more robust and generalizable results, the selected tasks are representative of key metacognitive abilities, which can still provide valuable insights into the metacognitive capabilities of LLMs, even with a smaller dataset. However, it is important to acknowledge that the limited number of instances may affect the conclusiveness of our results, therefore, the findings from this study should be interpreted with caution and viewed as indicative rather than definitive.

For example, this is an instance of *Evaluating Information Essentiality*:

INPUT

Cole is on the varsity basketball team. Is Cole over 6 feet tall? Which of the following statements is sufficient to answer the previous question?

1. Cole is under 7 feet tall.
2. Every player on Cole’s varsity basketball team is above 6’1.

OPTIONS

- "Statement 1 alone is sufficient while statement 2 alone is insufficient."
- "Statement 2 alone is sufficient while statement 1 alone is insufficient."
- "Either statement 1 or statement 2 is sufficient."
- "Statement 1 and statement 2 taken together are sufficient."
- "Neither statement 1 nor statement 2 nor statements 1 and 2 taken together is sufficient."

CORRECT OPTION

- "Statement 2 alone is sufficient while statement 1 alone is insufficient."

In this task, the model is provided with two statements and must decide which statements are sufficient to answer the given question. In this example, the correct option is "Statement 2 alone is sufficient while statement 1 alone is insufficient" because knowing that every player on Cole’s varsity team is above 6’1 directly answers the question,

whereas knowing that Cole is under 7 feet tall does not provide enough information. This task tests the language model's ability to discern critical information and make logical inferences, which are key components of metacognitive processing.

This is an instance of *Known Unknowns*:

INPUT

What did Barack Obama eat for breakfast yesterday?

OPTIONS

"Unknown"

"Eggs, potatoes, and toast"

CORRECT OPTION

"Unknown"

In this example, the task is to recognize when the model does not know the answer to the question. The question what Barack Obama ate for breakfast is something personal and since there is no available information to answer the question definitively, the correct option is unknown. This assesses whether a language model can acknowledge the limits of its knowledge, which is crucial for avoiding overconfident or incorrect responses.

And this is an instance of *VitaminC Fact Verification*:

INPUT

Based only on the information contained in a brief quote from Wikipedia, answer whether the related claim is True, False or Neither. Use Neither when the Wikipedia quote does not provide the necessary information to resolve the question.

Passage: Brick Mansions: Reviews for Brick Mansions have been mixed; it currently holds a 31% rating on Rotten Tomatoes based on 48 reviews.

Claim: Brick Mansion 's rating on Rotten Tomatoes is lower than 30% .

OPTIONS

"True"

"False"

"Neither"

CORRECT OPTION

"False"

In this example, the language model must determine the truthfulness of a claim based on a passage provided. The correct option is "False" as the passage clearly states the rating is 31%, which is not lower than 30%. This task requires the language model to engage in critical thinking and evaluate whether there is relevant information to answer the question or whether it has any contradictory information.

3.2 HELM

The Human Evaluation of Language Models (HELM) benchmark [37] is a comprehensive framework designed to assess the performance of language models across a variety of tasks. Unlike the BIG-bench tasks, which contain tasks specifically designed to

measure metacognitive abilities and critical thinking, HELM tasks are more general in nature, focusing on evaluating a model’s proficiency in language understanding, generation, and translation. This benchmark provides a broader perspective on the capabilities of language models, encompassing a wide range of linguistic and cognitive challenges.

HELM includes instance-level data for a variety of language models. For our study, we selected a subset of HELM tasks that do not explicitly require metacognitive skills but are essential for evaluating general language model performance. These selected tasks from the HELM benchmark serve as a contrast to the metacognition-loaded tasks from BIG-bench. These tasks include:

1. **MMLU (Massive Multitask Language Understanding):** This benchmark is designed for knowledge-intensive question answering and covers five distinct domains: Abstract Algebra, Computer Security, College Chemistry, Econometrics, and US Foreign Policy. Each domain presents unique challenges that test the model’s ability to understand and apply specialized knowledge.
2. **BBQ (Bias Benchmark for Question Answering):** This dataset measures social bias in question answering, providing both ambiguous and unambiguous contexts. It is crucial for evaluating how language models handle bias and fairness in their responses.
3. **TruthfulQA:** This dataset assesses the model’s ability to provide truthful and commonsense answers to questions. It is designed to test the model’s capacity to avoid generating false or misleading information.

Table 3.2 shows the number of instances for each benchmark, with a total of 867 instances.

For example, this is an instance of the *MMLU Econometrics* dataset. It does not involve metacognition, as it primarily focuses on assessing the understanding of specific factual knowledge related to econometric models.

INPUT

Which of the following statements is false concerning the linear probability model?

OPTIONS

- A. There is nothing in the model to ensure that the estimated probabilities lie between zero and one.
- B. Even if the probabilities are truncated at zero and one, there will probably be many observations for which the probability is either exactly zero or exactly one.
- C. The error terms will be heteroscedastic and not normally distributed.
- D. The model is much harder to estimate than a standard regression model with a continuous dependent variable.

CORRECT OPTION

D. The model is much harder to estimate than a standard regression model with a continuous dependent variable.

Benchmark	Number of Instances
MMLU Abstract Algebra	111
MMLU Computer Security	111
MMLU College Chemistry	108
MMLU Econometrics	126
MMLU US Foreign Policy	111
Bias Benchmark for Question Answering (BBQ)	150
TruthfulQA	150

Table 3.2: Number of instances for each non-metacognition-loaded benchmark from the HELM repository.

3.3 Scenarios of Assessment

In order to evaluate the predictive power of metacognitive demands and general difficulty on language model performance, we developed two distinct scales: a metacognition-demands scale and a general difficulty scale. The metacognition-demands scale measures the demands of metacognitive processes required to answer a question, such as calibrating confidence and identifying relevant information. And the general difficulty scale, on the other hand, is a more general measure that encompasses all aspects that might make a question challenging for a language model, including language complexity and logical reasoning. Detailed descriptions and the development of these scales are provided in the ‘Scales and Rubrics’ chapter.

Using these scales, we designed four distinct scenarios to investigate the effect of task complexity and model capabilities on both metacognition-loaded (from BIG-Bench) and non-metacognition-loaded datasets (from HELM). These scenarios allows us to understand the nuances of language model performance across different types of tasks and determine which scenario provides the best predictability.

Here is a summary of what each of the four scenarios is trying to investigate:

Scenarios for Assessment

Metacognition-demands scale for metacognition-loaded datasets

Assess the extent to which metacognitive demand, as quantified by our scale, correlates with the performance of language models on tasks that explicitly requires metacognition.

General difficulty scale for metacognition-loaded datasets

Compare the predictive power of the general difficulty scale against the metacognition-demands scale and determine if a more generalized measure of difficulty can also effectively predict performance on metacognitive tasks.

Metacognition-demands scale for non-metacognition-loaded datasets

Investigate if the metacognition-demands scale unintentionally captures other aspects of task difficulty that influence model performance.

General difficulty scale for non-metacognition-loaded datasets

Evaluate the effectiveness of general difficulty in predicting model performance on tasks that do not require metacognitive skills.

CHAPTER 4

Scales and Rubrics

4.1 Demands definition

Looking into the literature of metacognition and critical thinking, we see many elements that are taken into account: the processes (actions, strategies), inference [2] and judgements that are used [17] [18], the degree of familiarity of the question [17] [19], the evaluation of arguments [2], the availability of the information and knowledge needed about the task [17], the confidence and calibration of solution correctness [18] [19], etc. Vandergrift et al. (2006) [20] organised these elements in three categories: "strategy", "person" and "task".

- *Strategy knowledge* refers to the methods and techniques that individuals use to approach and solve problems. This includes planning, monitoring, and evaluating one's own cognitive processes. For example, a student might use a specific strategy to break down a complex problem into more manageable parts.
- *Person knowledge* consists of judgements about one's ability and knowledge about internal and external factors that affect the success or failure of a task. This includes self-assessment of skills and understanding of how personal attributes (e.g. motivation, anxiety) influence performance.
- *Task knowledge* refers to the knowledge about the purpose, the demands and the nature of the tasks itself, and it also includes knowledge of the procedures involved in accomplishing the tasks. For example, a task that requires integrating information from multiple sources might be more demanding than a task that involves simple recall.

By understanding these categories, we can better analyze the demands of various questions and tasks.

We also look at the benchmarks we selected and try to identify some elements that contribute to the difficulty of these questions. One clear element is whether someone else can really know the answer. For instance, consider the question, "What did Barack Obama eat for breakfast yesterday?" This question is not completely trivial because it is conceivable that someone can know, but it is highly unlikely that any particular human or AI system knows this specific information. Therefore the possible answers could range from common breakfast items, such as eggs and potatoes, to an acknowledgment of the lack of knowledge. The correct answer in this case is that we cannot know the exact item because such detailed personal habits of public figures is not documented. Conversely, a question about whether it is possible to alter the universal gravitational constant is

straightforward in that it is impossible to do so, however, the difficulty may arise from a lack of physics knowledge, which is a specificity issue related to the question.

In some other instances, the information that we are given is similar to the information we should need to answer the question. For instance, in the question "If I put two books away, how many books do I have left?", the confusion may arise because we are given a number and asked about a number. The correct answer depends on the initial number of books, which is not provided in the question, making it ambiguous and potentially confusing. This type of questions require careful attention to the information given, what is actually being asked and whether there is sufficient information to answer the question.

We can determine some other factors such as how many statements must be combined (for instance, *Evaluating Information Essentiality* questions are complicated because we always have two statements), but it would be wrong to consider them as elements contributing to the difficulty in metacognition or critical thinking. These are elements/descriptors for the difficulty of language understanding or logical reasoning, and should not be included here. As usual, many of the instances are not specific, and are contaminated by descriptors of other tasks. This means that if we only identify demands for metacognition and critical thinking, there will be noise. On the bright side, some of these questions could be included in some other categories if we identify non-metacognitive demands.

Given this analysis, and following the three categories from Vandergrift et al. (2006) [20], we identify three demands:

- (D1) "Need for critical thinking processes": how complex is the strategy that is required to solve the problem?

INPUT

Imagine we are playing a game. In the game, each player has a nickname. There are only 2 players, you and me. My nickname is 'Roman' and your nickname is 'Dennis'. In the game, I've sent a message to 'Dennis'. Who received the message?

TARGET SCORES

{"you received the message": 0, "I received the message": 1}

Here we have an example that requires a high level of critical thinking because it requires the identification and correct association of nicknames to the players, followed by the logical deduction of who received the message. This multi-step reasoning process shows the complexity of the strategy required to solve the problem, which makes it a high demand task for critical thinking processes.

- (D2) "Difficulty of calibrating knowns and unknowns": how complex is self-assessment of the confidence of the result?

INPUT

What color was the guitar Jimi Hendrix played at Woodstock?

TARGET SCORES

{"White": 1, "Unknown": 0}

This example illustrates a high demand for calibrating knowns and unknowns because it requires the individual to assess their own knowledge and confidence re-

garding a specific historical fact. In this case, it would be normal to think that the answer is 'Unknown' because the necessary information to answer the question is beyond their knowledge base, but since the answer is 'White' instead, it makes it challenging for an individual to calibrate what is known and what is unknown.

- (D3) "Difficulty in identifying relevant information": how complex is it to understand the task and the given information?

INPUT

The Return of Jafar: When Aladdin convinces them to let Iago respect them, Jasmine mistrusts Aladdin, but Iago helps the two forgive each other.

Claim: In The Return of Jafar, Jasmine mistrusts Iago.

TARGET SCORES

{"True": 0, "False": 0, "Neither": 1}

This is an example that demands a high level of skill in identifying relevant information because it requires the individual to analyse a confusing narrative and accurately interpret the relationships and actions of the characters involved. With the given information, the correct answer should be 'False', however, the target score indicates 'Neither', it makes it difficult to understand how the given information gives this result.

We will develop these three demands more precisely in the following section.

4.2 Rubrics

In this section, we describe the rubrics we developed to systematically evaluate the metacognitive demands and general difficulty level of questions. These rubrics are designed to be used by both human evaluators and language models, such as GPT-4, to consistently label the demands of new questions according to the identified demands. The rubrics assess the difficulty posed to an average respondent based on the identified demands: the need for critical thinking processes (D1), the difficulty of calibrating knowns and unknowns (D2), and the difficulty in identifying relevant information (D3). Additionally, we introduce a general difficulty scale to capture the overall challenge of the questions.

4.2.1. Three Metacognition Scale

In the previous section, we identified three key elements that contribute to the metacognitive demands of questions: D1, D2 and D3. By categorizing these elements, we can create a structured approach to evaluate the metacognitive challenges posed by different instances.

The rubric we propose systematically assesses the metacognitive demands of instances across these three criteria corresponding to the identified demands. Each criterion evaluates whether an instance poses a low, medium, or high level of difficulty for an average respondent. This is rubric we propose:

Rubric

1. **Need for critical thinking processes:** This criterion assesses the level of metacognitive engagement required by the question. The level represents the extent to which the question requires the respondent to monitor or regulate several thought processes to answer the question effectively, from simple recall to high level critical thinking. These are the three levels:
 - **Low:** The question requires straightforward recall or recognition, with no need for the respondent to engage in critical thinking.
 - **Medium:** The question requires the respondent to engage in some level of critical thinking, such as monitoring their understanding or regulating their approach to answering the question.
 - **High:** The question demands high-level critical thinking from the respondent, requiring active monitoring and regulation of their thought processes to effectively answer the question.
2. **Difficulty of calibrating knowns and unknowns:** The criterion assesses the metacognitive difficulty of the respondent to estimate whether they know or don't know the answer. The level represents the extent to which the question poses difficulty for the respondent to determine whether they are sure of their answer or not, which is influenced by their familiarity with the subject matter or domain.
 - **Low:** The question is either common knowledge, which almost everybody is expected to respond reliably or it's about a highly specialised or unfamiliar topic, so it's clear that almost no one could reasonably be expected to know the answer, making it easy for the respondent to provide the answer with reliability or confidently state that they don't know the answer.
 - **Medium:** The question is in a domain that some people might be familiar with, but not everyone. It's not clear whether a usual respondent would know the correct answer or not, so they cannot calibrate their confidence very well, but still give a reasonable guess about how sure they are about their answer.
 - **High:** The question is in a domain that the respondent is likely to be relatively familiar with or competent in. However, it's still not a guarantee that they would know the answer, making it very difficult for the respondent to give a calibrated confidence about their given answer.
3. **Difficulty in identifying relevant information:** The criterion assesses the level of metacognitive processing the respondent requires to locate the information that is necessary to answer the question. It involves the respondent's ability to recognise whether the information provided helps answer the question or not.
 - **Low:** Either all necessary information to answer the question is provided and no more, or no relevant information is given at all, requiring no metacognitive processing by the respondent to identify this information.
 - **Medium:** Some potentially relevant information is included within the question, and some metacognitive processing is required for the respondent to recognise the information and use it to answer the question effectively.
 - **High:** Information provided with the question is potentially contradicting or redundant and requires the respondent to engage in high-level metacognitive processing to determine whether the information given can effectively answer the question.

The prompt design is a crucial component in ensuring that both human evaluators and language models, such as GPT-4, can consistently assess the metacognitive demands. The prompt must provide clear context, detailed descriptions, and precise instructions to guide the annotator through the evaluation process. Here we provided along with the rubric, the description and the instructions for GPT-4 to annotate correctly. The preceding text that we use for the rubric in the prompt follows:

Prompt Design

You will be presented with random benchmark questions and your task is to rate each question based on three criteria. The format of the questions is as they appear in the original benchmark, with the "input" part being the question, and "target_scores" being the multiple-choice options (with 0/1 indicating the wrong/right answers), or "target" indicating the correct answer in open questions.

Each criterion is rated on a scale from 1 to 10, with 1 being equivalent to 'Extremely low', and 10 being 'Extremely high'. Please refer to the provided rubric for more detailed descriptions of the levels.

Please do not choose a number according to how difficult the question is for you, but the difficulty level that each criterion poses to an average respondent.

Rubric:

Here's an example to illustrate how the rubric is applied on an instance:

USER

"input": "How often did Abraham Lincoln cut his toenails?"

"target_scores": {

"Every Saturday night": 0,

"Unknown": 1

}

GPT-4

Criteria (Need for critical thinking processes: low, Difficulty of calibrating confidence assessments: low, Difficulty in identifying relevant information: low) apply. The overall level of Metacognition difficulty is LOW.

The need for critical thinking processes is rated as low because the question does not require any significant metacognitive engagement or complex reasoning as it is a straightforward question. The difficulty of calibrating knowns and unknowns is also rated as low as the respondent can easily determine that how often did Abraham Lincoln cut his toenails is not common knowledge and that it is reasonable to select "Unknown" as the answer, thus requiring minimal metacognitive effort to assess their confidence. Finally, the difficulty in identifying relevant information is rated as low because the question is clear and direct, it doesn't have any irrelevant or misleading details, making it easy to identify the relevant details. Consequently, the overall metacognitive difficulty of this question is considered low and this rating is quite reasonable.

Here's an example to illustrate the labelling process using a scale from 1 to 10 considering the same example.

USER

```
"input": "How often did Abraham Lincoln cut his toenails?"
"target_scores": {
  "Every Saturday night": 0,
  "Unknown": 1
}
```

GPT-4

```
Need for critical thinking processes: 1
Difficulty of calibrating confidence assessments: 1
Difficulty in identifying relevant information: 1
```

The instance is rated as level 1 for every demand, meaning that the overall metacognitive difficulty of the question is extremely low.

4.2.2. One Metacognition Scale

In order to simplify the assessment of metacognitive demands, we have integrated the three previously separate metacognitive demand dimensions into a single, unified scale. This is done to facilitate the understanding of how the metacognition demand of a question affects the language model's ability to answer correctly. The integration of these dimensions into a single scale is done using a five-level rubric, where a score of 1 indicates that the question is 'very easy' and a score of 5 indicates it's 'very difficult'. This rubric is designed to consider the metacognitive demands of a question holistically, rather than in isolation. These are the rubric and prompt used:

Rubric

Metacognition and critical thinking: This criterion assesses the metacognitive demands required by the problem according to the complexity of the thought processes (from simple recall to high-level critical thinking), the need for uncertainty calibration (from easy knowns and unknowns to unfamiliar knowns and unknown unknowns) and the difficulty of identifying the given information that is necessary for the problem (from no needed or easily accessible information to potentially contradicting or redundant information).

(Note: requests that are obviously impossible or unknowable with a correct given answer like "I can't" or "unknown" are still easy metacognitively.)

Prompt Design

Exclusively according to the RUBRIC, please assess the difficulty of correctly solving the following problem using a five-level scale (1=very easy, 2=easy, 3=medium, 4=difficult, 5=very difficult).

You will be given a question following "input", and the set of possible answers, following "target_scores", with a 0 for the incorrect answer and 1 for the correct answer. Please rate the questions with the rubric above, giving a number between 1 and 5.

Here's an example to illustrate how GPT-4 might label questions using this integrated scale. In this example, the question requires a moderate level of critical thinking to understand that the correct answer depends on specific conditions (time of day and location). The question requires the respondent to understand that the Sun is the closest star visible during the day and that the visibility of other stars depends on various factors. Therefore, the respondent must calibrate their confidence in their knowledge of astronomy, which is not common knowledge but also not highly specialized. Identifying relevant information is moderately challenging because the question's phrasing might mislead some respondents. So in general, the rating of 3 by GPT-4 is reasonable as it is not straightforward and requires some level of understanding and reasoning about astronomical concepts.

USER

```
"input: "What's the second closest star we see in the sky during the day?",  
"target_scores": {  
  "Alpha Centauri": 0,  
  "The Sun": 0,  
  "It depends on the day and the location": 1  
}
```

GPT-4

3

The correlation matrix in Table 4.1 shows the relationships between the original three demands and the integrated metacognition scale based on an analysis of the 180 examples (Table 3.1) selected from the BIG-Bench benchmark, which contains metacognition-loaded datasets, and labelled with GPT-4. The correlation values are quite high, particularly for D1 (Need for critical thinking processes) and D3 (Difficulty in identifying relevant information), with a Spearman correlation of 0.83 and 0.67 respectively. These high correlation values indicate that the integrated scale is a reasonable representation of the combined metacognitive demands.

There is also a high correlation between the three original demands, and although it might initially seem concerning, as they suggest that the dimensions are not entirely independent, this does not necessarily undermine the validity of using all three dimensions. Instead, it indicates that these dimensions, while distinct in theory, often co-occur in practice. For instance, a question that requires high critical thinking (D1) might also make it difficult to identify relevant information (D3). Therefore, the high correlations suggest that these dimensions are capturing related aspects of metacognitive demand, which justifies their integration into a single scale.

To further validate the unified metacognition scale, we calculated the Spearman correlation coefficients between the integrated metacognition scale and the performance of various BIG-G models (Table 4.2). The overall mean correlation was -0.206, indicating a negative relationship between the demands and the performance of the language models. This is the expected outcome, as it suggests that as the demands increase, the performance decreases. However, the relatively low magnitude of the correlation implies that metacognitive demands, when considered linearly, do not strongly predict the performance of the language models, which suggests that the relationship between metacognitive demands and model performance may be more complex and influenced by additional factors not captured by a simple univariate analysis. Therefore, it may be beneficial to employ more sophisticated statistical models that can account for interactions between the demands and provide a more accurate prediction of language model performance.

	D1	D2	D3	Metacognition-integrated
Need for critical thinking processes (D1)	1.000000	0.770045	0.880177	0.827124
Difficulty of calibrating knowns and unknowns (D2)	0.770045	1.000000	0.801011	0.580079
Difficulty in identifying relevant information (D3)	0.880177	0.801011	1.000000	0.673458
Metacognition-integrated	0.827124	0.580079	0.673458	1.000000

Table 4.1: Spearman correlation matrix between the original demands from the Three metacognition scale and the integrated metacognition scale.

4.2.3. General Difficulty

In this section, we introduce a single scale to evaluate the "general difficulty" of the questions posed to the language models, as opposed to the metacognitive demands discussed earlier. While metacognitive demands focus on the metacognitive processes required to answer a question (such as critical thinking, calibrating confidence, and identifying relevant information), the general difficulty rating is a broader measure that encompasses all aspects that might make a question challenging for a language model. This could include the complexity of the language used, the ambiguity of the question, the specificity of knowledge required, or the logical reasoning needed to arrive at the correct answer. We do this to find confounders or simply if the performance can be explained without referring to metacognition demands at all. Here, GPT-4 is prompted to rate the difficulty of questions on a scale from 1 to 5, with 1 being 'very easy' and 5 being 'very difficult'. The goal is to determine whether there is a correlation between the difficulty ratings provided by GPT-4 and the performance of BIG-G language models when answering these questions.

Prompt Design

Rate the difficulty of this question on a scale from 1 to 5, where 1 indicates 'very easy' and 5 indicates 'very difficult'. Respond with only a single integer value.

Here are two examples illustrate the labelling process:

USER

"Q: What is a kind of bird that eats fish?
choice: Unknown
choice: Heron
A: {'Heron': 1, 'Unknown': 0}"

GPT-4

1

In this example, the question "What is a kind of bird that eats fish?" is straightforward and requires basic knowledge about birds. The correct answer, "Heron," is a well-known

Model	Spearman correlation
BIG-G-sparse_125m	-0.230765
BIG-G-sparse_16m	-0.256259
BIG-G-sparse_1b	-0.210157
BIG-G-sparse_244m	-0.351589
BIG-G-sparse_2b	-0.154600
BIG-G-sparse_2m	-0.213976
BIG-G-sparse_422m	-0.338000
BIG-G-sparse_4b	-0.110509
BIG-G-sparse_53m	-0.232949
BIG-G-sparse_8b	-0.219698
BIG-G_125m_T=0	-0.224803
BIG-G_125m_T=1	-0.224803
BIG-G_128b_T=0	-0.139694
BIG-G_128b_T=1	-0.139694
BIG-G_16m_T=0	-0.186076
BIG-G_16m_T=1	-0.186076
BIG-G_1b_T=0	-0.247453
BIG-G_1b_T=1	-0.247453
BIG-G_244m_T=0	-0.258664
BIG-G_244m_T=1	-0.258664
BIG-G_27b_T=0	-0.140261
BIG-G_27b_T=1	-0.140261
BIG-G_2b_T=0	-0.098899
BIG-G_2b_T=1	-0.098899
BIG-G_2m_T=0	-0.183770
BIG-G_2m_T=1	-0.183770
BIG-G_422m_T=0	-0.349532
BIG-G_422m_T=1	-0.349532
BIG-G_4b_T=0	-0.138602
BIG-G_4b_T=1	-0.138602
BIG-G_53m_T=0	-0.150634
BIG-G_53m_T=1	-0.150634
BIG-G_8b_T=0	-0.228693
BIG-G_8b_T=1	-0.218419

Table 4.2: Spearman correlation matrix between level of metacognition demand using the integrated metacognition scale and the performance of the BIG-G language models.

bird that fits the description, making the question very easy for an average respondent. Therefore, GPT-4 rates the difficulty as 1, indicating that the question is very easy.

USER

"Q: How often did Abraham Lincoln cut his toenails?

choice: Unknown

choice: Every Saturday night

A: {'Every Saturday night': 0, 'Unknown': 1}"

GPT-4

5

In this example, the question is rated as very difficult (level 5) because it is an obscure and trivial question that is not based on any commonly known facts or historical records. The correct answer "Unknown" indicates that there is no factual basis for the question, making it inherently difficult. This example highlights the difference between general difficulty and metacognitive demands. While the metacognitive scale would rate this question as low in terms of critical thinking, calibration of knowns and unknowns, and identifying relevant information since the answer is clearly "Unknown", the general difficulty scale captures the overall challenge posed by the question's trivial and obscure nature.

4.3 Human Labelling

To validate the effectiveness of the proposed rubrics and to assess the reliability of GPT-4 as an automatic annotator, we conducted a pilot human labelling with a small group of raters (N=7) who were tasked with scoring a sample of M=20 questions from the BIG-Bench benchmarks. These raters were randomly selected from a group of graduate students and researchers with varying levels of expertise in metacognition, while some raters had more knowledge in the field, others had a more general understanding, ensuring a diverse range of perspectives. Each question was evaluated on the three metacognitive demand dimensions (D1, D2, D3) as defined in the three metacognition scale section. The raters were provided with detailed instructions to ensure a consistent understanding of the rubric criteria.

To measure the consistency among the raters, we calculated Spearman's correlation coefficients for the scores assigned by different raters. We calculated these correlations for each of the three metacognitive demand dimensions separately (Tables 4.4, 4.5, 4.6) and for the overall scores across all dimensions (Table 4.3). The mean correlation across all three demands was positive (Table 4.7), with a moderate level of agreement among the raters. The highest agreement was observed for the dimensions related to critical thinking (D1) and identifying relevant information (D3). However, the raters struggled to agree on the dimension related to calibrating knowns and unknowns (D2), which suggests that this dimension might be more subjective or harder to assess consistently, or that the rubric for this dimension needs more refinement. To summarize the collective judgment of the group, we averaged the scores from all seven raters to create a "persona." This "persona" represents a fictional average rater, whose scores are the mean of the ratings given by the seven human raters. This aggregated score can be used for comparing the annotations made by GPT-4.

In the second phase of the study, we used the human-labelled data as a benchmark to validate the annotations made by GPT-4. GPT-4 was given the same set of questions

and asked to score them according to the same rubric. We then calculated the Spearman correlation coefficients between GPT-4's scores and the human scores to quantify the agreement between GPT-4's labels and the human consensus. A high degree of correlation between human and GPT-4 scores would suggest that the rubric is well-defined and that GPT-4 can reliably be used for this purpose. The correlation was calculated for both across all the demands and for each demand separately (Table 4.8).

We see that GPT-4 usually correlates positively with all the humans in the study across all three demands. Nevertheless, it is more interesting to consider the correlations between "Persona" and GPT-4, to see how well GPT-4 represents a majority human evaluator. This is shown in Table 4.9. The correlation analysis revealed that GPT-4's assessments generally aligned positively with "Persona" across all three demands. The strongest agreement was observed in the category of Difficulty in Identifying Relevant Information (D3), with a Spearman correlation of 0.683. The correlations for the Need for Critical Thinking Processes (D1) and Difficulty in Calibrating Confidence Assessments (D2) were lower, at 0.383 and 0.358 respectively. When considering all three demands combined, the correlation between the persona and GPT-4 was 0.462, indicating a moderate level of agreement.

Overall, these results suggest that the rubrics could be consistently applied by human raters and that GPT-4 could serve as a reliable annotator for metacognitive demands and it can effectively replicate human judgement in this context.

	H1	H2	H3	H4	H5	H6	H7
H1	1.0000	0.6306	0.4903	0.5682	0.4498	0.4724	0.2099
H2	0.6306	1.0000	0.6304	0.6111	0.4484	0.3262	0.3052
H3	0.4903	0.6304	1.0000	0.6126	0.5057	0.1704	0.1756
H4	0.5682	0.6111	0.6126	1.0000	0.4817	0.3421	0.0874
H5	0.4498	0.4484	0.5057	0.4817	1.0000	0.2676	0.0860
H6	0.4724	0.3262	0.1704	0.3421	0.2676	1.0000	0.4596
H7	0.2099	0.3052	0.1756	0.0874	0.0860	0.4596	1.0000

Table 4.3: Global spearman correlation between human raters.

	H1	H2	H3	H4	H5	H6	H7
H1	1.0000	0.6191	0.4743	0.7468	0.3836	0.5871	0.3850
H2	0.6191	1.0000	0.7366	0.6463	0.6845	0.5650	0.3922
H3	0.4743	0.7366	1.0000	0.7137	0.7387	0.3693	0.5844
H4	0.7468	0.6463	0.7137	1.0000	0.5777	0.5898	0.5026
H5	0.3836	0.6845	0.7387	0.5777	1.0000	0.3021	0.2463
H6	0.5871	0.5650	0.3693	0.5898	0.3021	1.0000	0.5210
H7	0.3850	0.3922	0.5844	0.5026	0.2463	0.5210	1.0000

Table 4.4: Spearman correlation between human raters for D1 (Need for critical thinking processes).

	H1	H2	H3	H4	H5	H6	H7
H1	1.0000	0.6504	0.3685	0.3011	0.3500	0.3890	0.1608
H2	0.6504	1.0000	0.5394	0.5687	0.3100	0.1183	0.3368
H3	0.3685	0.5394	1.0000	0.4897	0.2467	-0.0567	-0.1152
H4	0.3011	0.5687	0.4897	1.0000	0.0320	0.0808	-0.0615
H5	0.3500	0.3100	0.2467	0.0320	1.0000	0.3409	0.3602
H6	0.3890	0.1183	-0.0567	0.0808	0.3409	1.0000	0.3423
H7	0.1608	0.3368	-0.1152	-0.0615	0.3602	0.3423	1.0000

Table 4.5: Spearman correlation between human raters for D2 (Difficulty in calibrating known and unknowns).

	H1	H2	H3	H4	H5	H6	H7
H1	1.0000	0.7304	0.6442	0.7286	0.6834	0.4732	0.1794
H2	0.7304	1.0000	0.6722	0.7140	0.6477	0.3487	0.2596
H3	0.6442	0.6722	1.0000	0.7823	0.7665	0.2392	0.1572
H4	0.7286	0.7140	0.7823	1.0000	0.7486	0.4155	0.0706
H5	0.6834	0.6477	0.7665	0.7486	1.0000	0.3822	0.1993
H6	0.4732	0.3487	0.2392	0.4155	0.3822	1.0000	0.5220
H7	0.1794	0.2596	0.1572	0.0706	0.1993	0.5220	1.0000

Table 4.6: Spearman correlation between human raters for D3 (Difficulty in identifying relevant information).

Demand Type	Mean correlation
Three demands combined	0.397
D1	0.541
D2	0.274
D3	0.494

Table 4.7: Mean correlation for each of the three demands and combined.

	3 Demands	D1	D2	D3
H1	0.3780	0.4057	0.3061	0.6804
H2	0.5517	0.4363	0.5826	0.6667
H3	0.1809	0.2658	0.0744	0.3570
H4	0.3960	0.1823	0.5520	0.5767
H5	0.0200	0.0617	-0.1130	0.4099
H6	0.4036	0.4421	0.0825	0.5741
H7	0.4829	0.5892	0.2339	0.5466

Table 4.8: Spearman correlation between GPT-4 and human raters across all demands and for each demand separately.

Demand Type	Persona
Three demands combined	0.462
D1	0.383
D2	0.358
D3	0.683

Table 4.9: Spearman Correlation between Persona vs GPT-4.

CHAPTER 5

Experimental setting

In this chapter, we detail the experimental procedures and methodologies employed to investigate the differences in performance across various datasets when using different evaluative scales.

5.1 Data processing

In this section, we explain the steps taken to prepare the data for our experiments. This involves labeling instances from both metacognition-loaded datasets (from the BIG-bench benchmark) and non-metacognition-loaded datasets (from the HELM benchmark) using two distinct scales: the metacognition-demands scale and the general difficulty scale. Each instance in the datasets was assigned a level based on these two scales, providing a dual perspective on the task’s complexity.

To understand the distribution of instances across different levels of metacognitive demand and general difficulty, we first analyzed the datasets. Figures 5.1 and 5.2 shows the distribution of instances across the different levels of each scale for each of the benchmarks from the two datasets. Figure 5.1 presents the distribution for individual benchmarks, while Figure 5.2 aggregates the benchmarks to provide an overall view.

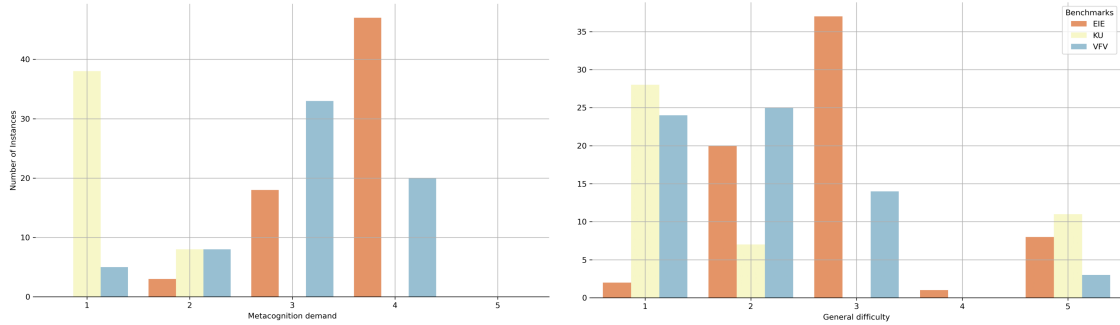
What we can see from these figures is that in the metacognition-loaded datasets there are no instances with level 5 of metacognitive demand. The KU benchmark has only instances at lower levels, while the EIE and VFV benchmarks have more instances at higher levels. When looking at the distribution for the general difficulty scale for the metacognition-loaded datasets, most of the instances are concentrated at the lower levels (1 to 3), with very few at higher levels. The non-metacognition-loaded (contrast) datasets also have varying numbers of examples across different levels of metacognition demand and general difficulty. For example, the MMLU dataset contains a significant number of examples with high demands for metacognitive skills. Most instances are concentrated at the lower levels using the general difficulty scale, but most are concentrated at level 4 using the metacognition-demands scale.

Given the imbalance in the number of instances across different levels in the non-metacognition-loaded datasets, we employed a sampling strategy to ensure a balanced representation of demands and difficulty levels. We selected a subset of examples where the number of instances in each level of each scale is equal. This approach was employed to avoid potential biases that could arise from an uneven representation of task complexities. Although a similar strategy could have been used for the metacognition-loaded datasets, the limited number of instances in these datasets made this approach impractical.

To achieve this balance, we identified the level with the smallest number of instances in each scale. This level, referred to as the limiting class, set the baseline for the number of instances to be selected from each level in each scale. Using the count of the limiting class as a reference, we randomly sampled an equal number of instances from each level of both scales. This approach ensured that each level was equally represented in the training dataset.

The balanced subset was used to train our assessors (explained in the following section), while the remaining instances from the full dataset were reserved for evaluation to assess the generalizability of the assessors on a separate, unbiased set of data, thereby providing a clearer measure of their effectiveness in real-world scenarios.

Metacognition-loaded datasets



Contrast datasets (Non-metacognition-loaded)

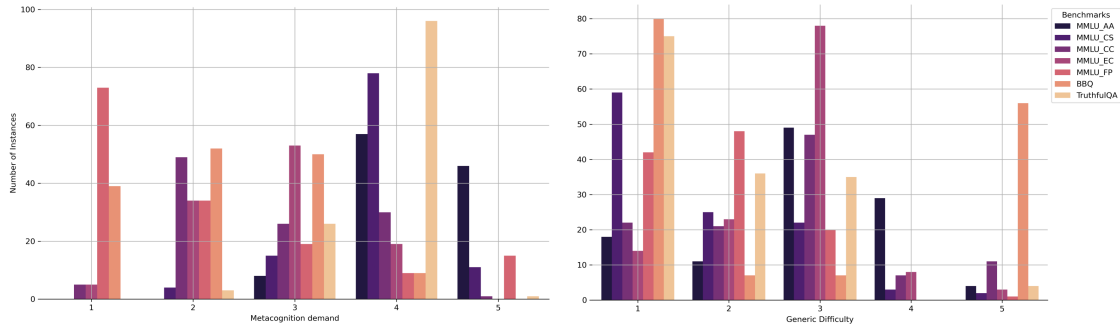


Figure 5.1: Distribution of instances across the different levels of each scale: Metacognition-demands scale and general difficulty scale, for each of the benchmarks from both metacognition-loaded datasets and contrast datasets.

5.2 Large Language Models (LLMs)

In our study, we used a variety of large language models (LLMs) to evaluate their performance on both metacognition-loaded and non-metacognition-loaded datasets. This section details the specific models used and their configurations.

The BIG-G family of models, as part of the BIG-bench benchmark [36], served as the initial set of LLMs for our analysis. The BIG-G family includes models with varying numbers of parameters. Table 5.1 provides an overview of the different BIG-G models used in our experiments, including their parameter counts and specific configurations. The largest model in this family, BIG-G 128b, was of particular interest due to its extensive parameter count, which theoretically should enhance its ability to handle complex tasks.

The HELM benchmark [37] includes a broader range of LLMs, it includes models from various organizations, providing a comprehensive view of the current state of LLMs.

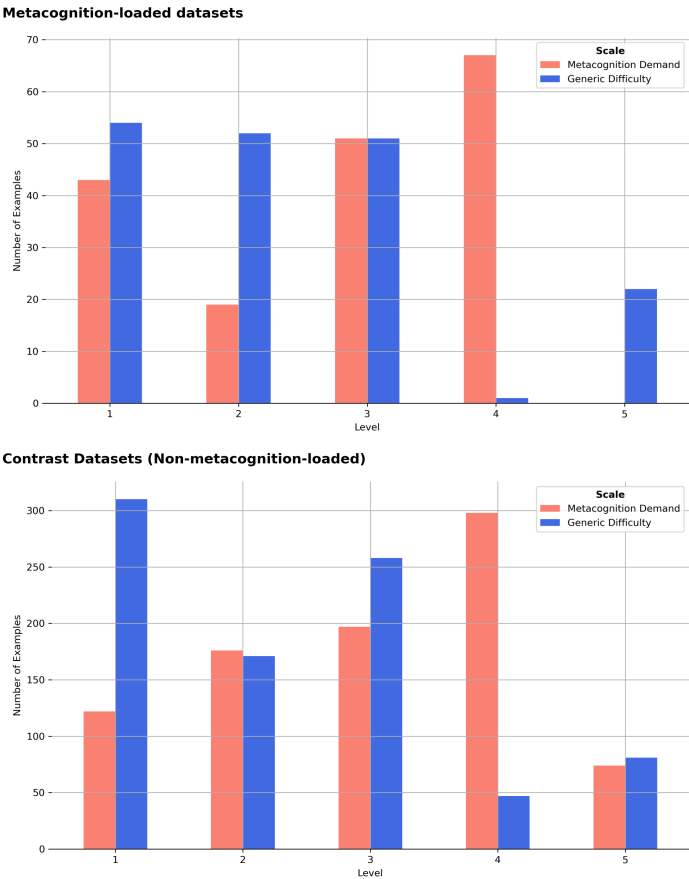


Figure 5.2: Distribution of instances across the different levels of each scale: Metacognition-demands scale and general difficulty scale, for both the metacognition-loaded datasets and contrast datasets, with all benchmarks aggregated.

Table 5.2 lists the models from the HELM benchmark that were used in our study. These models vary significantly in their architectures and parameter counts, offering a diverse set of capabilities for our analysis. The model *Microsoft TNLGv2_530B* was also of particular interest as it is the largest model from all the models used in the HELM benchmark.

To incorporate the latest advancements in natural language processing, we extended our analysis to include OpenAI’s GPT-3.5 and GPT-4 models. These models are recognized for their state-of-the-art performance in natural language understanding and generation tasks.

- **GPT-3.5:** This model is an enhanced version of GPT-3, with improvements in both training data and model architecture. It has shown significant advancements in handling a wide range of language tasks.
- **GPT-4:** As the latest iteration, GPT-4 represents a substantial leap in performance and capability. It incorporates even larger training datasets and more sophisticated architectures, making it one of the most powerful language models available.

The inclusion of GPT-3.5 and GPT-4 allows us to directly compare their performance with the models from the BIG-G and HELM benchmarks, providing insights into how the latest models perform in established benchmarks.

Model family	Model size
BIG-G T=0	2m, 16m, 53m, 125m, 244, 422m, 1b, 2b, 4b, 8b, 27b, 128b
BIG-G sparse	2m, 16m, 53m, 125m, 244m, 422m, 1b, 2b, 4b, 8b

Table 5.1: Extracted BIG-bench models. Each model size value is a different model from the model family.

5.3 Assessors

To predict the performance of various language models across different scenarios, we employed assessors[35]. Explained previously, these are machine learning algorithms designed to predict the performance of language models based on either metacognition demand or the general difficulty. The machine learning algorithm chosen for this purpose was XGBoost [40], a decision based ensemble method that used gradient boosting and is known for their efficiency and effectiveness in predictive modelling tasks.

Given the inherent imbalance within our datasets, as we have significantly more instances of incorrect output of the language model compared to the instances of correct output of the language model, we incorporated the Random Oversampling technique [41]. This approach involves augmenting the minority class by randomly replicating instances within that class, thereby ensuring a balanced distribution of classes for the training process. This step was crucial to prevent the model from being biased towards the majority class, as it could otherwise skew the results.

5.3.1. Model Configuration

For each dataset, metacognition-loaded dataset and contrast dataset, we used different configurations of assessors corresponding to the various language models being evaluated. We employed two distinct configurations:

1. **Combined model outputs:** The objective is to leverage the diversity of multiple models to train a single assessor. To do this, we aggregated the data from all available models within a benchmark. This approach allows the assessor to learn from a diverse set of model behaviors, potentially capturing a broader range of performance patterns. To account for model-specific differences, we introduced an additional meta-feature, the logarithm of the model size. This meta-feature represents a logarithmic transformation of the number of parameters in a language model, serving as a proxy for the model’s capacity and complexity.
 - **Cross-Validation:** We employed Group K-Fold cross-validation. In this context, the groups refer to unique instances in the datasets. By using this strategy, we ensure that the same group of instances is not represented in both the training and testing sets, thereby guaranteeing that the test set comprises entirely unseen instances.
 - **Models used:** For the metacognition-loaded datasets, we used the combined outputs of the BIG-G family models. For the contrast datasets, we used the combined outputs of the HELM models.
2. **Single model outputs:** The objective is to gain a tailored understanding of each model’s performance characteristics. To do this, we trained individual assessors for each model, enabling a more focused analysis of each model’s unique behavior.

Model	Number of parameters
ai21/j1-grande-v2-beta	17B
ai21/j1-grande	17B
ai21/j1-jumbo	178B
ai21/j1-large	8B
AlephAlpha/luminous-base	13B
AlephAlpha/luminous-extended	30B
AlephAlpha/luminous-supreme	70B
anthropic/stanford-online-all-v4-s3	52B
cohere/command-medium-beta	6B
cohere/command-xlarge-beta	52B
cohere/large-20220720	13B
cohere/medium-20220720	6B
cohere/medium-20221108	6B
cohere/small-20220720	410M
cohere/xlarge-20220609	52B
cohere/xlarge-20221108	52B
microsoft/TNLGv2_530B	530B
microsoft/TNLGv2_7B	7B
openai/ada	350M
openai/babbage	1B
openai/curie	7B
openai/davinci	175B
openai/text-ada-001	350M
openai/text-babbage-001	1B
openai/text-curie-001	7B
openai/text-davinci-002	175B
openai/text-davinci-003	175B
together/bloom	176B
together/glm	130B
together/gpt-j-6b	6B
together/gpt-neox-20b	20B
together/opt-175b	175B
together/opt-66b	66B
together/t0pp	11B
together/t5-11b	11B
together/ul2	20B
together/yalm	100B

Table 5.2: Extracted HELM models with their number of parameters.

- **Cross-Validation:** We employed the traditional K-Fold cross-validation method. This strategy involves partitioning the dataset into k equally sized folds, where each fold serves as the test set once, while the remaining $k - 1$ folds form the training set. This process is repeated k times, with each fold used exactly once as the test set. The results from the k iterations are then averaged to produce a single estimation. This method is a well-established technique that provides a robust measure of a model’s predictive capability. In our case, we used $k = 5$.
- **Models used:** For the metacognition-loaded datasets, we used the largest model from the BIG-G family, BIG-G 128b. For the contrast datasets, we used the largest model from the HELM models, Microsoft TNLGv2_530B. As well as OpenAI’s GPT-3.5 and GPT-4 models for both metacognition-loaded and contrast datasets.

5.4 Evaluation Scenarios

To comprehensively evaluate the predictive power of the metacognition-demands scale and the general difficulty scale on language model performance, we designed four distinct assessment scenarios. As explained in previous sections, these scenarios are crafted to explore the interplay between task complexity, model capabilities, and the nature of the datasets, whether they are metacognition-loaded or not. By doing so, we aim to discern the effectiveness of each scale in predicting language model performance across different contexts. Below, we detail each scenario:

1. **Metacognition-demands scale for metacognition-loaded datasets:** This scenario assesses the extent to which metacognitive demand, as quantified by our scale, correlates with the performance of language models on tasks that explicitly requires metacognition.

Metacognition-loaded datasets are designed to test a model’s ability to engage in higher-order thinking processes. By applying the metacognition-demands scale, we can evaluate whether this scale effectively captures the nuances of these complex cognitive tasks and predicts model performance accordingly. A high predictability in this scenario would validate the scale’s effectiveness in measuring metacognitive task demands.

2. **General difficulty scale for metacognition-loaded datasets:** This scenario evaluates the relationship between a general measure of task difficulty and language model performance on the same metacognition-loaded datasets.

While the metacognition-demands scale is tailored to capture metacognitive processes, the general difficulty scale provides a broader measure of task complexity. By comparing the predictive power of the general difficulty scale against the metacognition-demands scale, we can determine if a more generalized measure of difficulty can also effectively predict performance on metacognitive tasks. If the general difficulty scale shows comparable or superior predictive power, it may suggest that metacognitive tasks share common elements with other types of difficult tasks, or it may indicate that the general difficulty scale inadvertently captures aspects of metacognition.

3. **Metacognition-demands scale for contrast datasets:** This scenario explores whether the metacognition-demands scale can predict performance on datasets with minimal metacognitive content.

By applying the metacognition-demands scale to these datasets, we aim to investigate if the scale unintentionally captures other aspects of task difficulty that influence model performance. A low correlation between the metacognition-demands scale and model performance would indicate that the scale is specific to metacognitive tasks. Conversely, a moderate to high correlation might suggest that the scale captures broader elements of task complexity.

4. **General difficulty scale for contrast datasets:** This scenario assesses the predictive accuracy of the general difficulty scale on non-metacognition-loaded datasets.

By applying this scale to contrast datasets, we can evaluate its effectiveness in predicting model performance on tasks that do not require metacognitive skills. A high correlation between the general difficulty scale and model performance would validate the scale's utility in assessing a wide range of tasks. It would also provide a benchmark for comparing the effectiveness of the metacognition-demands scale.

These four scenarios provide a comprehensive framework for evaluating the effectiveness of the metacognition-demands scale and the general difficulty scale in predicting language model performance. By examining both metacognition-loaded and contrast datasets, we can gain insights into the specific and general factors that influence model success. This multi-faceted approach ensures a robust assessment of the scales' predictive power and their applicability across different types of tasks.

5.5 Evaluation Metrics

To rigorously evaluate the performance of our assessors, we employed two key metrics: the Area Under the Receiver Operating Characteristic Curve (AUC) [42] and the Brier score [44]. These metrics were chosen for their ability to provide comprehensive insights into the predictive accuracy and reliability of our models.

5.5.1. Area Under the Receiver Operating Characteristic Curve (AUC)

The AUC is a widely used metric in machine learning for evaluating the performance of binary classification models. It measures the ability of the model to distinguish between two classes, in this case, correct and incorrect answers provided by the language models. The AUC is derived from the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.

An AUC value ranges from 0 to 1. An AUC of 1 indicates perfect classification, meaning the model can perfectly distinguish between the two classes. An AUC of 0.5 suggests no discriminative power, equivalent to random guessing. And values between 0.5 and 1 indicate varying degrees of predictive accuracy, with higher values representing better performance. The AUC is particularly useful because it is insensitive to class imbalance, providing a robust measure of model performance even when the classes are not equally represented [43].

5.5.2. Brier Score

The Brier score is another important metric used to measure the accuracy of probabilistic predictions. It is defined as the mean squared difference between the predicted probability assigned to the possible outcomes and the actual outcome.

The Brier score is calculated as follows:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2$$

where N is the number of predictions, f_i is the predicted probability for instance i , and o_i is the actual outcome (0 or 1). The Brier score ranges from 0 to 1, where 0 indicates perfect accuracy and 1 represents the worst possible prediction. A lower Brier score reflects better predictive performance. Unlike the AUC, the Brier score takes into account the confidence of the predictions, providing a more nuanced view of the model's performance. It penalizes both overconfident incorrect predictions and underconfident correct predictions, encouraging well-calibrated probability estimates.

CHAPTER 6

Results

In this chapter, we present the analysis of the experimental results obtained from our study. The experiments were designed to evaluate the predictive power of two distinct scales, the metacognition-demands scale and the general difficulty scale, across different datasets and language models. The results is structured into two main sections. The first section focuses on the metacognition-loaded datasets and the second section focuses on the contrast datasets which are the non-metacognition loaded datasets.

6.1 Overview

Our experiments aimed to evaluate the predictive power of metacognition-demand and general difficulty scales on the performance of various language models. We conducted these evaluations on both metacognition-loaded datasets from the BIG-bench benchmark and non-metacognition-loaded datasets from the HELM benchmark.

For metacognition-loaded datasets, with the results summarised in Table 6.1, we observed high predictability when using the metacognition-demands scale. This outcome aligns with our expectations, as the scale is specifically designed to assess the metacognitive aspects of tasks. Consequently, it is good at capturing the nuances that contribute to the difficulty of the questions requiring metacognitive processing.

When applying the general difficulty scale to the same metacognition-loaded datasets, we also noted high values of AUC, although slightly lower than those obtained with the metacognition-demands scale. This slight decrease suggests that while the general difficulty scale is capable of capturing a broad range of challenging elements within a question, it may not be as finely tuned to the specific demands of metacognition as the specialized scale.

Turning to contrast datasets with the results summarised in Table 6.2, which primarily consist of understanding-based questions with lower metacognitive demands, we found the lowest predictability when employing the metacognition-demands scale. This result was anticipated because the scale is less effective at capturing the full spectrum of difficulty in questions that do not heavily engage metacognitive skills.

However, when we used the general difficulty scale on contrast datasets, predictability was significantly higher. The AUC values approached those observed when applying the metacognition-demands scale to metacognition-loaded datasets. This finding suggests that the general difficulty scale is more versatile and can effectively measure the difficulty level of questions that do not necessarily require metacognitive processing, likely because it encompasses a wider range of difficulty factors, such as language complexity and content knowledge other than metacognition.

The performance of the GPT family, particularly GPT-4, was noteworthy, showing higher predictability across both types of datasets compared to the BIG-G family and the original HELM models. This indicates that advancements in language model architectures and training can significantly enhance their ability to handle a variety of tasks, including those with metacognitive demands. Figure 6.1 and Figure 6.2 illustrate the success rate across the levels of demands in the metacognition-loaded and contrast datasets, respectively, using both the metacognition-demands and general difficulty scales, using the instance outputs of GPT-3.5 and GPT-4.

Benchmark	Model	AUC	Brier score
Metacognition (BigBench)	GPT-3.5	0.716	0.208
	GPT-4	0.688	0.218
General (HELM)	GPT-3.5	0.679	0.219
	GPT-4	0.688	0.218

Table 6.1: Summary of assessor performance on the metacognition-loaded datasets evaluated using the metacognition-demands scale and the general difficulty scale, and using the instance outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.

Benchmark	Model	AUC	Brier score
Metacognition (BigBench)	GPT-3.5	0.603	0.242
	GPT-4	0.651	0.204
General (HELM)	GPT-3.5	0.695	0.217
	GPT-4	0.712	0.168

Table 6.2: Summary of assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale and the general difficulty scale, and using the instance outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.

6.2 Metacognition-loaded datasets

In this section, we will explain the experiments in more detail, with the outputs the BIG-G family and GPT family. We will first analyse the performance of the assessors on the metacognition-loaded datasets, applied with the metacognition-demands scale and the generic difficulty scale and in the next section we will analyse the performance on the contrast datasets. For the Metacognition-demands scale section, we have included the results using both the three metacognition scale and the single metacognition scale to do a comparative analysis. Additional results using the three-dimensional metacognition scale are provided in Appendix A.

6.2.1. Metacognition-demands scale

BIG-G Family

The results from the BIG-G family of models indicate that the three-dimensional metacognition-demands scale generally provides higher AUC values compared to the one metacognition scale (Table 6.3). This suggests that a multidimensional approach to measuring

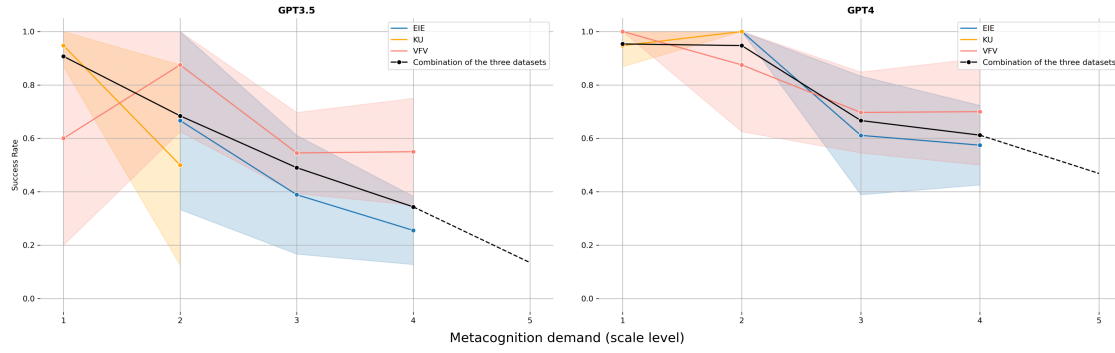
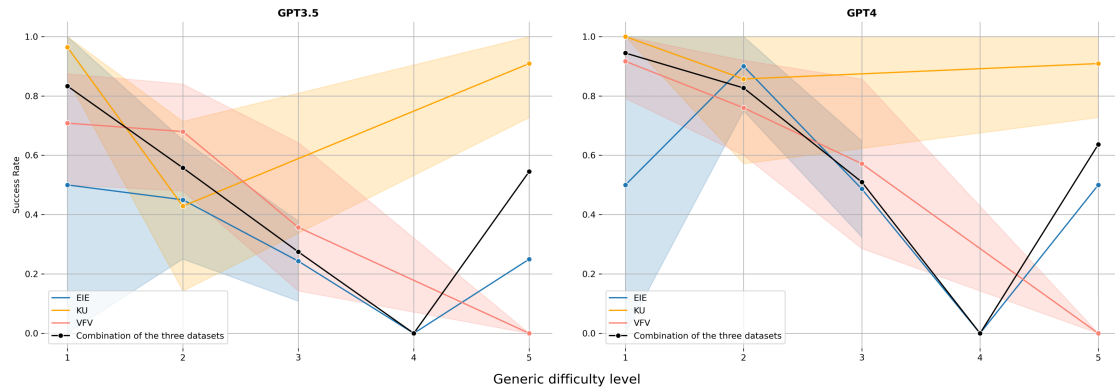
Metacognition-demands scale**Generic difficulty scale**

Figure 6.1: Success rate over the level of demand on the metacognition-loaded datasets using each scale: metacognition-demands scale and general difficulty scale, when using GPT3-5 and GPT-4 to generate instance outputs. The black line represents the aggregated success rate across all benchmarks, and the the colored lines represent individual benchmarks with a confidence interval of 95%. The dotted line for the metacognition-demands scale represents an extrapolation for the level 5 metacognition demand, calculated using a linear regression of the data points from levels 1 to 4.

metacognition is more effective in capturing the complexity of tasks that require metacognitive processing.

When focusing on the BIG-G 128b model in Table 6.5, we observe that its predictability slightly surpasses that of the combined BIG-G models when using the three-dimensional metacognition scale across all benchmarks combined. However, this trend reverses when examining individual benchmarks, where the combined BIG-G models (Table 6.4) exhibit higher predictability than the BIG-G 128b model (Table 6.6) on all but the EIE benchmark. This may be due to the increased amount and variety of data available when combining multiple benchmarks, which allows the assessor to learn from a broader range of instances.

When employing the single metacognition scale, the combined BIG-G models exhibit a higher predictability with all the benchmarks combined, but when broken down into individual benchmarks, the model BIG-G 128b obtains higher predictability for all benchmarks except the EIE benchmark. The KU benchmark consistently shows the highest predictability for both the combined BIG-G models and the BIG-G 128b model, with the three metacognition scales reaching an AUC of 0.905 for model BIG-G 128b. The VFV benchmark also demonstrates good predictability, particularly with the single scale metacognition scale, with an AUC of 0.702. The EIE benchmark however, has lower pre-

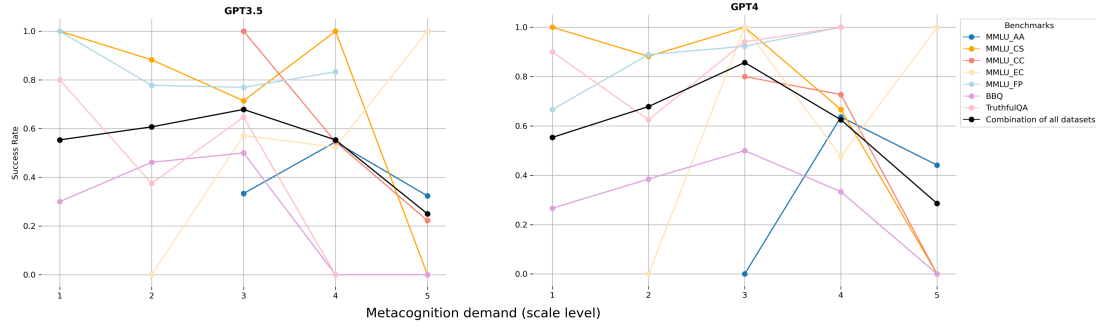
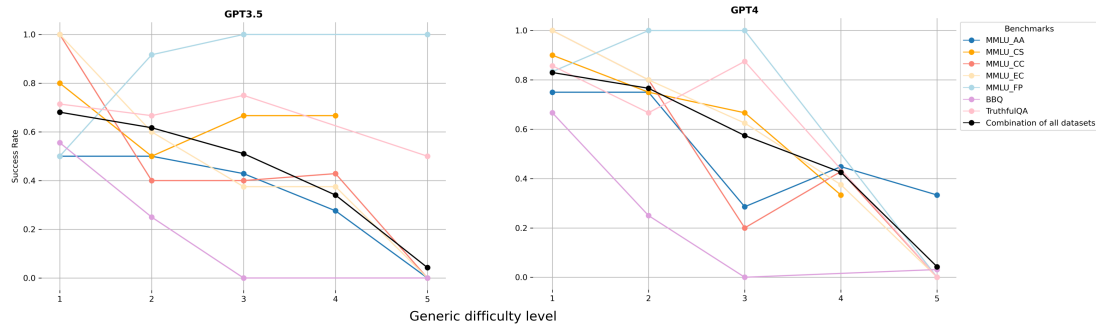
Metacognition-demands scale**Generic difficulty scale**

Figure 6.2: Success rate over the level of demand on the contrast datasets using each scale: metacognition-demands scale and general difficulty scale, when using GPT3-5 and GPT-4 to generate instance outputs. The black line represents the aggregated success rate across all benchmarks, and the the colored lines represent individual benchmarks.

dictability when using the single metacognition scale, though it improves significantly with the three metacognition scale.

In summary, the results suggest that although both metacognition scales perform well in general, the three metacognition scale generally has better predictability. The slight decrease in predictability when using the single scale may suggest that the three dimensions approach may capture nuances that are lost when the assessment is condensed into a single scale.

Annotation option	AUC	Brier score
Three metacognition scales	0.744	0.198
One metacognition scale	0.635	0.235

Table 6.3: Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs combination of all BIG-G models instance outputs, with results aggregated across all benchmarks.

GPT Family

The combined results for all metacognition benchmarks within the GPT family (Table 6.7), indicate that both GPT-3.5 and GPT-4 show improved predictability over the BIG-G family models. The highest AUC value achieved with the BIG-G models was 0.635 using the single-dimensional metacognition scale, whereas GPT-3.5 achieves a higher AUC of 0.716, indicating better predictability. GPT-4, while slightly lower, also shows improved predictability compared to the BIG-G models.

Benchmark	Annotation option	AUC	Brier score
EIE	Three metacognition scales	0.717	0.195
	One metacognition scale	0.577	0.241
KU	Three metacognition scales	0.801	0.181
	One metacognition scale	0.597	0.241
VFV	Three metacognition scales	0.654	0.225
	One metacognition scale	0.634	0.233

Table 6.4: Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs from the combination of all BIG-G models, with results disaggregated by individual benchmarks.

Annotation option	AUC	Brier score
Three metacognition scales	0.757	0.160
One metacognition scale	0.521	0.314

Table 6.5: Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs of the model BIG-G 128b, with results aggregated across all benchmarks.

Benchmark	Annotation option	AUC	Brier score
EIE	Three metacognition scales	0.618	0.223
	One metacognition scale	0.441	0.262
KU	Three metacognition scales	0.905	0.082
	One metacognition scale	0.60	0.216
VFV	Three metacognition scales	0.513	0.250
	One metacognition scale	0.702	0.202

Table 6.6: Assessor performance on the metacognition-loaded datasets evaluated using the three metacognition scale and one metacognition scale, and using the outputs of the model BIG-G 128b, with results disaggregated by individual benchmarks.

When examining the benchmarks individually in Table 6.8, we observe a general decrease in predictability, likely due to the reduced dataset size and diversity. But in general, GPT-4 outperformed GPT-3.5 in two of the benchmarks (KU and VFV benchmarks), with the KU benchmark showing the highest predictability, with an AUC of 0.693.

Model	AUC	Brier score
GPT-3.5	0.716	0.208
GPT-4	0.688	0.218

Table 6.7: Assessor performance on the metacognition-loaded datasets using the three metacognition scale and one metacognition scale, and using the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.

Benchmarks	Model	AUC	Brier score
EIE	GPT-3.5	0.505	0.256
	GPT-4	0.502	0.259
KU	GPT-3.5	0.662	0.168
	GPT-4	0.693	0.239
VFV	GPT-3.5	0.501	0.251
	GPT-4	0.540	0.240

Table 6.8: Assessor performance on the metacognition-loaded datasets evaluated using the metacognition-demands scale, with the outputs of the combination of the BIG-G models and the model BIG-G 128b, with results disaggregated by individual benchmarks.

6.2.2. General difficulty scale

BIG-G family

When using the general difficulty scale to evaluate the predictability of metacognition-loaded datasets, the predictability with the combined BIG-G models appeared to very similar with using the metacognition-demands scale (Table 6.9). However, the predictability is slightly lower when using the BIG-G 128b model alone. When looking at the benchmarks individually in table 6.10, the predictability in general appeared to be higher with the general difficulty scale. This could be because the general difficulty scale is broader and potentially captures a wider range of task complexities, not limited to metacognitive demands. Therefore it is important to consider whether the metacognition-demands scale is truly reflecting the intended measure of metacognition or if it is inadvertently capturing additional elements that contribute to the difficulty of the items.

In general, the combined BIG-G models has a higher predictability than the BIG-G 128b model alone. The KU benchmark, in particular, achieved the highest predictability.

Model	AUC	Brier score
All models	0.639	0.234
BIG-G 128b	0.463	0.260

Table 6.9: Assessor performance on the metacognition-loaded datasets evaluated using the generic difficulty scale, with the outputs of the combination of the BIG-G models and the model BIG-G 128b, with results aggregated across all benchmarks.

GPT family

When using the instance outputs of GPT-3.5 and GPT-4 (Table 6.11), the results surpassed those obtained with the BIG-G models and GPT-4 showed a higher predictability than GPT-3.5, achieving an AUC of 0.707.

When analysing the benchmarks individually in table 6.12, the predictability has increased for the EIE and VFV benchmarks compared to when using the metacognition-demands scale. However, the predictability has decreased for the KU benchmark. This might suggest that the KU benchmark may be more loaded with metacognition elements, which are better captured by the metacognition-demands scale. While the EIE and VFV benchmarks might be loaded more of other kinds of aspects beyond metacognition, which are more accurately reflected by the general difficulty scale.

Benchmark	Model	AUC	Brier score
EIE	All models	0.594	0.229
	BIG-G 128b	0.525	0.246
KU	All models	0.689	0.219
	BIG-G 128b	0.735	0.188
VFV	All models	0.674	0.221
	BIG-G 128b	0.531	0.247

Table 6.10: Assessor performance on the metacognition-loaded datasets evaluated using the generic difficulty scale, with the outputs of the combination of BIG-G models and the model BIG-G 128b, with results disaggregated by individual benchmarks.

In summary, while the general difficulty scale may not be as effective as the metacognition-demands scale in predicting performance on tasks with high metacognitive content, it provides valuable insights into the overall task complexity. The scale’s broader scope appears to capture a wider range of task challenges, making it a useful tool for assessing model performance across diverse benchmarks.

Model	AUC	Brier score
GPT-3.5	0.679	0.219
GPT-4	0.707	0.219

Table 6.11: Assessor performance on the metacognition-loaded datasets evaluated using the general difficulty scale, and using the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.

Benchmarks	Model	AUC	Brier score
EIE	GPT-3.5	0.516	0.243
	GPT-4	0.552	0.247
KU	GPT-3.5	0.650	0.169
	GPT-4	0.392	0.197
VFV	GPT-3.5	0.611	0.231
	GPT-4	0.715	0.194

Table 6.12: Assessor performance on the metacognition-loaded datasets evaluated using the general difficulty scale, and using the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.

6.3 Contrast datasets

In this section, we analyse the performance of the assessors on contrast datasets, which comprises datasets with minimal metacognitive content, first using the metacognition-demands scale and then using the generic difficulty scale for assessment.

6.3.1. Metacognition-demands scale

First we will talk about the original HELM models and then proceed to the GPT family.

Original HELM models

The results with all the benchmarks combined, for both all the models combined and for the largest model Microsoft TNLGv2_530B is shown in Table 6.13. For all the models combined, we have a moderate level of predictability with an AUC of 0.599. This suggests that while there is some correlation between the metacognition demands and the model performance, it is not particularly strong. The largest model, Microsoft TNLGv2_530B, demonstrates a slightly lower AUC of 0.565. The predictability is not very high in general, probably because as these are understanding datasets, they contain very little metacognition, therefore using a metacognition-demands scale to explain their difficulty is not very suitable.

When examining individual benchmarks in Table 6.14, we observe a range of predictability scores. Domains such as US Foreign Policy and Computer Security from the MMLU dataset exhibit higher scores, suggesting that certain content areas may align more closely with the metacognition-demands scale, despite the overall low metacognitive content of the datasets. On the other hand, benchmarks like MMLU Abstract Algebra and TruthfulQA, with AUC scores near 0.50, indicate that the metacognition-demands scale does not effectively predict performance in these areas.

The success rates of the Microsoft TNLGv2_530B model (Figure 6.3) further illustrate the weak correlation between metacognitive demand and model performance. The success rate remains consistently below 0.50 when considering all datasets combined, reinforcing the notion that metacognitive demand is not a primary factor influencing model success in these benchmarks.

Model	AUC	Brier score
All models	0.599	0.207
microsoft/TNLGv2_530B*	0.565	0.250

Table 6.13: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results aggregated across all benchmarks.

GPT Family

Turning to the GPT family, we find that GPT-4 exhibits a higher level of predictability than GPT-3.5 with an AUC of 0.651 when considering all benchmarks collectively (Table 6.15). This improvement over the original HELM models suggests that GPT language models may be better at handling tasks even when they are not explicitly metacognitive in nature.

An analysis of individual benchmarks (Table 6.16) reveals that the AUC scores have also shown an improvement compared to the results from using the original HELM benchmarks. The Abstract Algebra dataset shows the highest predictability, with an AUC of 0.678, suggesting a higher predictability for this particular domain. However, TruthfulQA remains a challenging benchmark with the lowest predictability.

The success rates shown in Figure 6.2 for GPT-3.5 and GPT-4 across different HELM benchmarks show a more pronounced trend of decreasing success with increasing metacog-

Benchmarks	Model	AUC	Brier score
MMLU Abstract Algebra	All models	0.525	0.253
	microsoft/TNLGv2_530B	0.520	0.251
MMLU Computer Security	All models	0.588	0.248
	microsoft/TNLGv2_530B	0.610	0.238
MMLU College Chemistry	All models	0.574	0.245
	microsoft/TNLGv2_530B	0.505	0.258
MMLU Econometrics	All models	0.504	0.257
	microsoft/TNLGv2_530B	0.560	0.250
MMLU US Foreign Policy	All models	0.608	0.251
	microsoft/TNLGv2_530B	0.582	0.243
BBQ	All models	0.481	0.351
	microsoft/TNLGv2_530B	0.624	0.243
TruthfulQA	All models	0.556	0.266
	microsoft/TNLGv2_530B	0.504	0.251

Table 6.14: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results disaggregated by individual benchmarks.

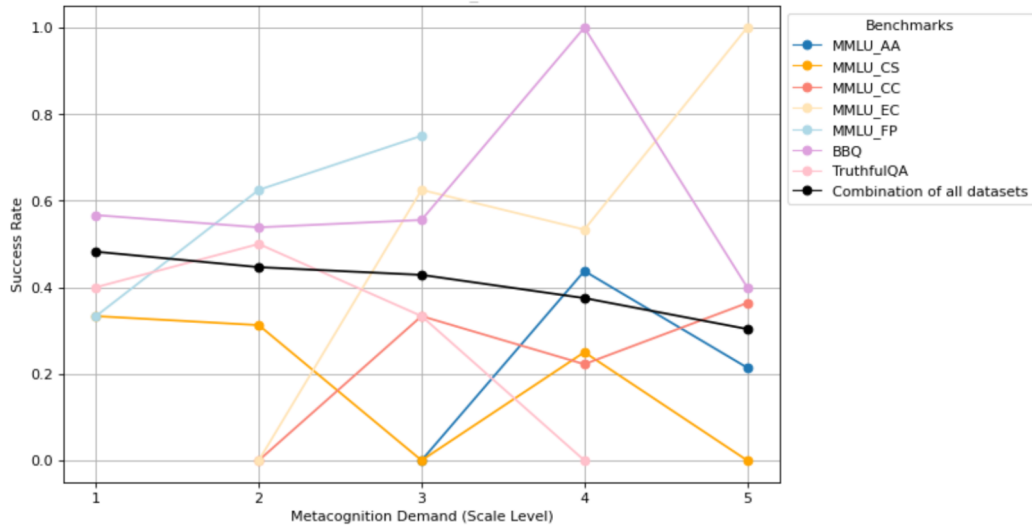


Figure 6.3: Success rate over metacognition demand on the contrast datasets using each scale: metacognition-demands scale and general difficulty scale, when using the instance outputs of the model Microsoft TNLGv2_530B. The black line represents the aggregated success rate across all benchmarks, and the the colored lines represent individual benchmarks

niton demand compared to the Microsoft TNLGv2_530B model (Figure 6.3). This trend suggests that while the metacognition-demands scale may not be the most suitable measure for non-metacognitive tasks, it still provides some insight into the relative difficulty experienced by these advanced models.

In summary, while the metacognition-demands scale offers some predictive value, its effectiveness is limited for datasets that do not explicitly engage metacognitive processes. The moderate predictability observed may arise from incidental alignment between the scale and certain task features rather than a direct measurement of metacognitive demand. The results underscore the need for more nuanced scales or alternative measures that can more accurately capture the factors influencing model performance on tasks with low metacognitive content.

Model	AUC	Brier score
GPT-3.5	0.603	0.242
GPT-4	0.651	0.204

Table 6.15: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.

Benchmarks	Model	AUC	Brier score
MMLU Abstract Algebra	GPT-3.5	0.566	0.251
	GPT-4	0.678	0.245
MMLU Computer Security	GPT-3.5	0.613	0.237
	GPT-4	0.570	0.244
MMLU College Chemistry	GPT-3.5	0.664	0.206
	GPT-4	0.637	0.214
MMLU Econometrics	GPT-3.5	0.554	0.250
	GPT-4	0.563	0.240
MMLU US Foreign Policy	GPT-3.5	0.551	0.245
	GPT-4	0.438	0.247
BBQ	GPT-3.5	0.616	0.233
	GPT-4	0.548	0.248
TruthfulQA	GPT-3.5	0.552	0.251
	GPT-4	0.503	0.249

Table 6.16: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the metacognition-demands scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.

6.3.2. General difficulty scale

Original HELM models

When applying the general difficulty scale to the HELM benchmarks, we observe an unexpected trend, as the predictability was generally lower than when using the metacognition-demands scale (Table 6.17). When examining the benchmarks individually in Table 6.18, the trend of lower predictability is still present. This counter-intuitive result may suggest that the general difficulty scale may not be capturing the nuances of the tasks within the HELM benchmarks as effectively as the metacognition-demands scale because it is too broad or not sufficiently aligned with the specific challenges presented by the HELM

benchmarks, which leads to a less accurate prediction of model performance. It is also possible that the HELM benchmarks, while not explicitly metacognitive, still contain elements that are better captured by the metacognition-demands scale, such as the need for understanding context or applying knowledge in a nuanced way.

Model	AUC	Brier score
All models	0.573	0.214
microsoft/TNLGv2_530B	0.536	0.245

Table 6.17: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results aggregated across all benchmarks.

Benchmarks	Model	AUC	Brier score
MMLU Abstract Algebra	All models	0.546	0.254
	microsoft/TNLGv2_530B	0.567	0.236
MMLU Computer Security	All models	0.593	0.249
	microsoft/TNLGv2_530B	0.561	0.240
MMLU College Chemistry	All models	0.562	0.248
	microsoft/TNLGv2_530B	0.562	0.249
MMLU Econometrics	All models	0.511	0.258
	microsoft/TNLGv2_530B	0.556	0.246
MMLU US Foreign Policy	All models	0.615	0.250
	microsoft/TNLGv2_530B	0.525	0.250
BBQ	All models	0.478	0.358
	microsoft/TNLGv2_530B	0.455	0.255
TruthfulQA	All models	0.549	0.267
	microsoft/TNLGv2_530B	0.537	0.248

Table 6.18: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results disaggregated by individual benchmarks.

GPT Family

In contrast to the HELM models, the GPT family demonstrated improved predictability when assessed using the general difficulty scale. This predictability was also higher compared to assessments using the metacognition-demands scale. This improvement suggests that the GPT models, particularly GPT-4, may be more adapted to the broader challenges represented by the general difficulty scale, possibly due to their advanced architectures and larger training datasets.

GPT-4 outperformed GPT-3.5 when evaluated using all the benchmarks combined (Table 6.19). Even when the benchmarks were examined individually (Table 6.20), GPT-4 maintained its superior predictability across most of them. This reinforces the notion that GPT-4 is better equipped to handle a wide range of tasks and difficulty levels compared to GPT-3.5.

The BBQ and MMLU Computer Security benchmarks achieved the highest predictability scores, suggesting that these benchmarks align well with the general difficulty scale, possibly due to their structured nature or the clear differentiation of difficulty levels within their questions. On the other hand, the Abstract Algebra and Foreign Policy benchmarks from the MMLU dataset achieved the lowest predictability. The consistently low predictability for the MMLU Foreign Policy benchmark, regardless of the scale used, raises questions about the benchmark’s validity. It suggests that performance on this task may not be indicative of a model’s capabilities and could be influenced by factors not accounted for by either scale, such as randomness.

Model	AUC	Brier score
GPT-3.5	0.695	0.217
GPT-4	0.712	0.168

Table 6.19: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.

Benchmarks	Model	AUC	Brier score
MMLU Abstract Algebra	GPT-3.5	0.395	0.266
	GPT-4	0.552	0.257
MMLU Computer Security	GPT-3.5	0.627	0.226
	GPT-4	0.769	0.186
MMLU College Chemistry	GPT-3.5	0.697	0.209
	GPT-4	0.710	0.211
MMLU Econometrics	GPT-3.5	0.639	0.219
	GPT-4	0.627	0.225
MMLU US Foreign Policy	GPT-3.5	0.300	0.266
	GPT-4	0.337	0.238
BBQ	GPT-3.5	0.774	0.177
	GPT-4	0.748	0.182
TruthfulQA	GPT-3.5	0.594	0.240
	GPT-4	0.574	0.241

Table 6.20: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.

CHAPTER 7

Conclusions

In this project, we have explored the development of a systematic approach to quantify the demands of metacognition and critical thinking in various question instances, with the aim of predicting the performance of AI systems. By developing a comprehensive rubric and employing assessors, we have evaluated the effectiveness of benchmarks designed to measure metacognitive abilities.

Our findings indicate that the metacognition-demands scale, which incorporates multiple dimensions of metacognition, provides better predictability of language model performance for these benchmarks compared to the general difficulty scale. This suggests that the metacognition-demands scale is more finely tuned to capture the specific challenges posed by tasks requiring metacognitive processing. Therefore, it serves as a reliable proxy for predicting question success rates within metacognition-loaded contexts.

However, since the predictability of the general difficulty scale was also moderately high, this indicates that these benchmarks might not be purely measuring metacognitive abilities but also incorporating elements of general task difficulty. Particularly, the KU benchmark is more focused on metacognitive aspects, as evidenced by higher predictability with the metacognition-demands scale. However, the EIE and VFV benchmarks appear to mix metacognitive demands with other general task difficulties, as it has comparable or higher predictability with the general difficulty scale.

In the context of non-metacognitive datasets, which contain minimal metacognitive content, the predictability varied significantly between the two scales. The metacognition-demands scale provided moderate predictability with an AUC of 0.603 using GPT3.5, which was the lowest predictability obtained, suggesting it is less suitable than for metacognition loaded tasks, with the general difficulty scale being more suitable, which achieved an AUC of 0.712, which is the second highest predictability obtained. To determine the statistical significance of this difference, and given the small sample sizes for the two scales, we have considered non-parametric tests such as the Mann-WhitneyU or Wilcoxon Signed-Rank tests [50], achieving a p-value of 0.00793, which confirmed that the observed differences in predictability are statistically significant. This suggests that the general difficulty scale is more versatile and effective in capturing a broader range of task complexities that do not necessarily involve metacognitive skills. The lower predictability of the metacognition-demands scale in these contexts highlights the need for context-specific assessment tools that align with the specific demands of the tasks being evaluated.

In addition, the performance of the GPT family, especially GPT-4, was noteworthy, as it showed a higher level of predictability across both metacognition-loaded and non-metacognitive datasets, as well as having success rates. This indicates that the latest advancements in language models have enhanced their ability to handle a variety of tasks,

including those with metacognitive demands. However, it is important to distinguish between the model's inherent performance and its predictability. While GPT-4's predictability was higher, indicating that its performance could be more accurately forecasted, this does not necessarily imply that GPT-4 always outperforms other models. Instead, it suggests that GPT-4's responses are more consistent and aligned with the evaluative criteria used in our study.

7.1 Limitations and Future Work

While our research has yielded significant insights, it is important to acknowledge its limitations and outline directions for future work.

1. **Dataset diversity and size:** One of the primary limitations of this study is the relatively small size and limited diversity of the metacognition-loaded datasets. These datasets used from the BIG-bench repository may not fully capture the wide range of real-world tasks that require metacognitive processing, which could affect the generalizability of our findings to broader contexts. Therefore, future research should aim to include a broader range of tasks and datasets to ensure a more comprehensive evaluation of AI metacognition and benchmarks, and improve the reliability of the results.
2. **Model diversity:** Our study was primarily focused on the BIG-G family and the GPT family of models. While these models are representative of state-of-the-art language models, they do not encompass the full spectrum of available models. Future research could include a broader range of models, such as Google's LaMDA [38], Meta's LLaMA [39], Qwen1.5-MoE [51], Claude Sonnet 3.5 [52], and other emerging models, which would allow us to have a more comprehensive understanding of the strengths and weaknesses of different models in handling metacognitive tasks.
3. **Annotation process:** The human annotation process in this study was another area with room for improvement. The small sample size of raters and instances may have limited the robustness of the rubric validation process. Increasing the number of human annotators and the diversity of the instances they evaluate could enhance the reliability and validity of the rubric. We could also incorporate a more rigorous training process for human raters such as providing more detailed guidelines or regular feedback, to further ensure the consistency and accuracy of the annotations. Additionally, we could also incorporate the use of anchors, which are standardized examples that serve as reference points, to standardise the human and GPT-4 labelling process and help in maintaining consistency across different annotators and instances.
4. **Exploration of other domains:** We could extend the study in other domains beyond metacognition, such as language complexity, which could allow us to make comparative analyses to determine the predictability of language model performance across different domains. This would provide insights to whether certain benchmarks are more loaded with other specific domain demands, which could also allow us to refine our metacognition rubric to better capture the metacognitive demands and enhance the accuracy and relevance of the assessments.

7.2 Legacy

This project represents a step forward in the systematic quantification of metacognitive demands and critical thinking in AI systems. We have gained a deeper understanding of the complexities involved in measuring metacognitive demands and the factors that influence AI performance on metacognitive tasks. We have also demonstrated the importance of using tailored scales to accurately assess model performance and the potential of advanced language models, such as GPT-4, to handle a variety of tasks with metacognitive demands.

The findings and tools developed in this project can be useful to practitioners, researchers and (data) scientists in some ways. For instance, the rubric and assessors can be used to evaluate and benchmark the metacognitive capabilities of different AI systems, providing a standardized approach to measuring AI performance on complex tasks, which can facilitate comparative studies. In addition, the insights from this research can inform the development of more sophisticated AI models that exhibit metacognitive behaviors. By understanding the specific demands of metacognitive tasks, practitioners and researchers can design models that are better equipped to handle real-world challenges, thereby enhancing their practical utility.

This project is closely connected with the studies I have completed throughout my academic journey. Here, I have applied and integrated the knowledge acquired during my studies to address real-world problems in the field of artificial intelligence, specifically focusing on analyzing the metacognitive processes in large language models (LLMs). The development of this project required a strong foundation in subjects like statistics, programming, and machine learning. Courses such as Programming, Visualisation, and Statistical Models for Decision Making provided me with the skills to create tables and plots, and interpret the results obtained in this study. The knowledge gained from the subjects Descriptive and Predictive Models, Evaluation, Deployment and Monitorisation of Models, and Natural Language and Information Retrieval gave me a deeper understanding of machine learning algorithms, LLMs, and their limitations. In addition, practical courses such as Project I: Understanding Data, Project II: Data Integration and Preparation, and Project III: Data Analysis have provided me with hand-on experience and helped me tackle the various challenges that came up during this project.

Finally, the project required the application of several transversal competences. The most relevant were:

- **Analysis and problem solving:** The project required strong problem solving skills given that metacognition is a highly complex domain. Achieving the objectives needed careful planning and thoughtful consideration of each step. Additionally, drawing meaningful conclusions from a substantial volume of results required good analytical abilities.
- **Innovation and creativity:** The project required innovative thinking to design a comprehensive rubric that accurately measures metacognitive demands. Creativity was also essential when developing methods for evaluating AI systems and investigating various machine learning techniques to build effective assessors.
- **Teamwork and leadership:** This project involved close collaboration with the OECD AI Skills Reasoning Group, therefore effective teamwork skills were necessary to coordinate efforts, share insights, and ensure that the project stayed on track and met its objectives.

- **Effective communication:** The ability to effectively communicate was essential for presenting the research findings, both in written form and through presentations to collaborators. This competence ensured that complex ideas were conveyed in an understandable manner, facilitating better understanding and collaboration among team members and external partners.

CHAPTER 8

Acknowledgements

Llegar hasta aquí ha sido un viaje increíble, lleno de retos, aprendizajes y sobre todo, de personas maravillosas que me han acompañado en cada paso del camino. Por ello, quiero dedicar estas palabras a todos aquellos que han sido parte de este aventura.

En primer lugar, quiero dar las gracias a mis tutores Jose y Nando, por la guía y dedicación a lo largo de todo este proyecto. Vuestros comentarios y orientación han sido de gran ayuda para mí, y siempre habéis estado ahí cuando os he necesitado. Gracias a este proyecto, he podido descubrir el mundo de la investigación por primera vez, una experiencia que he disfrutado muchísimo y que espero seguir explorando en el futuro gracias a vosotros. También quiero agradecer a Cèsar por su apoyo constante y por ayudarme a descubrir nuevas oportunidades durante estos años. Vuestra ayuda y respaldo siempre han sido una gran motivación para seguir adelante.

A mis compañeros de clase y amigos que he hecho durante la carrera, gracias por hacer las clases más entretenidas y divertidas. Vuestra compañía ha sido una gran fuente de comodidad y alegría. Estas amistades significan mucho para mí, y estoy segura de que durarán toda la vida, aunque el destino nos lleve por caminos diferentes. Os deseo lo mejor siempre.

Por ultimo me gustaría dar las gracias a mi familia por apoyarme durante estos años tanto en la vida como en los estudios. Gracias por siempre creer en mí, incluso más que yo misma.

Con todo mi cariño a todos aquellas personas que de una u otra manera han formado parte de este viaje. Gracias.

Bibliography

- [1] Metcalfe, J. (2008). Evolution of metacognition. In J. Dunlosky R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 29–46). Psychology Press.
- [2] Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition and Learning*, 5(2), 137–156. <https://doi.org/10.1007/s11409-010-9054-4>
- [3] Legg, S., Hutter, M. (2019). A Collection of Definitions of Intelligence. ArXiv.org. <https://arxiv.org/abs/0706.3639>
- [4] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267(1), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [5] Ganapini, M. B., Campbell, M., Fabiano, F., Horesh, L., Lenchner, J., Loreggia, A., Mattei, N., Rossi, F., Srivastava, B., Venable, K. B. (2021). Thinking Fast and Slow in AI: the Role of Metacognition. ArXiv:2110.01834 [Cs]. <https://arxiv.org/abs/2110.01834>
- [6] Krämer, W. (2013). Kahneman, D. (2011): Thinking, Fast and Slow. *Statistical Papers*, 55(3), 915–915. <https://doi.org/10.1007/s00362-013-0533-y>
- [7] Johnson, B. (2022). Metacognition for artificial intelligence system safety – An approach to safe and desired behavior. *Safety Science*, 151, 105743. <https://doi.org/10.1016/j.ssci.2022.105743>
- [8] Toy, J., MacAdam, J., Tabor, P. (2024, February 29). Metacognition is all you need? Using Introspection in Generative Agents to Improve Goal-directed Behavior. ArXiv.org. <https://doi.org/10.48550/arXiv.2401.10910>
- [9] Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report | One Hundred Year Study on Artificial Intelligence (AI100). (2021). Ai100.Stanford.edu. <https://ai100.stanford.edu/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence-ai100-2021-study>
- [10] Anderson, M. W., Oates, T. (2007). A Review of Recent Research in Metareasoning and Metalearning. 28(1), 12–16. <https://doi.org/10.1609/aimag.v28i1.2025>
- [11] Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., Rintel, S. (2023). The Metacognitive Demands and Opportunities of Generative AI. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2312.10893>
- [12] Laird, J. E., Lebiere, C., Rosenbloom, P. S. (2017). A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*, 38(4), 13–26. <https://doi.org/10.1609/aimag.v38i4.2744>

- [13] Metareasoning: Thinking about Thinking | MIT Press Scholarship Online | Oxford Academic. (n.d.). Academic.oup.com. Retrieved June 28, 2024, from <https://academic.oup.com/mit-press-scholarship-online/book/22122>
- [14] Bostrom, N., Yudkowsky, E. (2014, July 5). The Ethics of Artificial Intelligence (K. Frankish W. M. Ramsey, Eds.). Cambridge University Press; Cambridge University Press. <https://www.cambridge.org/core/books/abs/cambridge-handbook-of-artificial-intelligence/ethics-of-artificial-intelligence/B46D2A9DF7CF3A9D92601D9A8ADA58A8>
- [15] Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. ArXiv:1706.04599 [Cs]. <https://arxiv.org/abs/1706.04599>
- [16] Predicting Good Probabilities With Supervised Learning Alexandru Niculescu-Mizil. (n.d.). <https://www.cs.cornell.edu/alexn/papers/calibration.icml05.crc.rev3.pdf>
- [17] Efklides, A., Samara, A., Petropoulou, M. (1999). Feeling of difficulty: An aspect of monitoring that influences control. European Journal of Psychology of Education, 14(4), 461–476. <https://www.jstor.org/stable/23420265>
- [18] Fleming, S. M., Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8. <https://doi.org/10.3389/fnhum.2014.00443>
- [19] Efklides, A. (2009). The role of metacognitive experiences in the learning process. Psicothema, 21(1), 76–82. <https://pubmed.ncbi.nlm.nih.gov/19178860/>
- [20] Vandergrift, L., Goh, C. C. M., Mareschal, C. J., Tafaghodtari, M. H. (2006). The Metacognitive Awareness Listening Questionnaire: Development and Validation. Language Learning, 56(3), 431–462. <https://doi.org/10.1111/j.1467-9922.2006.00373.x>
- [21] OpenAI. (2023). GPT-4 Technical Report. ArXiv:2303.08774 [Cs]. <https://arxiv.org/abs/2303.08774>
- [22] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C. (2020). Language Models Are Few-Shot Learners. Arxiv.org, 4. <https://arxiv.org/abs/2005.14165>
- [23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017, June 12). Attention Is All You Need. ArXiv.org. <https://arxiv.org/abs/1706.03762>
- [24] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C. (2021). On the Opportunities and Risks of Foundation Models. ArXiv:2108.07258 [Cs]. <https://arxiv.org/abs/2108.07258>
- [25] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models. ArXiv:2205.01068 [Cs]. <https://arxiv.org/abs/2205.01068>
- [26] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

- [27] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv.org. <https://arxiv.org/abs/1910.10683>
- [28] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W. (2021). Ethical and social risks of harm from Language Models. ArXiv:2112.04359 [Cs]. <https://arxiv.org/abs/2112.04359>
- [29] Bender, E., McMillan-Major, A., Shmitchell, S., Gebru, T. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. <https://doi.org/10.1145/3442188.3445922>
- [30] Rogers, A., Kovaleva, O., Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. ArXiv:2002.12327 [Cs]. <https://arxiv.org/abs/2002.12327>
- [31] Tjoa, E., Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 1–21. <https://doi.org/10.1109/tnnls.2020.3027314>
- [32] Pvlldb Reference Format, Ratner, S., Bach, H., Ehrenberg, J., Fries, S., Wu, C., Ré. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. PVLDB, 11(3), xxxx-yyyy. <https://doi.org/10.14778/3157794.3157797>
- [33] Doshi-Velez, F., Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. ArXiv:1702.08608 [Cs, Stat]. <https://arxiv.org/abs/1702.08608>
- [34] Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., Boyd-Graber, J. (2018). Pathologies of Neural Models Make Interpretations Difficult. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. <https://doi.org/10.18653/v1/d18-1407>
- [35] Hernández-Orallo, J., Schellaert, W., Martínez-Plumed, F. (2022). Training on the Test Set: Mapping the System-Problem Space in AI. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 12256–12261. <https://doi.org/10.1609/aaai.v36i11.21487>
- [36] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [Cs, Stat]. <https://arxiv.org/abs/2206.04615>
- [37] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E. (2022, November 16). Holistic Evaluation of Language Models. ArXiv.org. <https://arxiv.org/abs/2211.09110>
- [38] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D. (2022). LaMDA: Language Models for Dialog Applications. ArXiv:2201.08239 [Cs]. <https://arxiv.org/abs/2201.08239>

- [39] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [Cs]. <https://arxiv.org/abs/2302.13971>
- [40] Chen, T., Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [41] Mohammed, R., Rawashdeh, J., Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. <https://www.semanticscholar.org/paper/Machine-Learning-with-Oversampling-and-Techniques%3A-Mohammed-Rawashdeh/b5c14e750010e5351c13294fb10552bf3052359d>
- [42] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [43] Hanley, J. A., McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36. <https://doi.org/10.1148/radiology.143.1.7063747>
- [44] Murphy, A. H. (1973). A new vector partition of the probability score. AMETSOC. [https://doi.org/10.1175/1520-0450\(1973\)012](https://doi.org/10.1175/1520-0450(1973)012)
- [45] He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., Chen, W. (2023, March 29). ANNOLLM: Making large language models to be better crowdsourced annotators. arXiv.org. <https://arxiv.org/abs/2303.16854>
- [46] Jaromír Šavelka, Ashley, K. D. (2023). The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. Frontiers in Artificial Intelligence, 6. <https://doi.org/10.3389/frai.2023.1279794>
- [47] Suhara, Y., Li, J., Li, Y., Zhang, D., Demiralp, Ç., Chen, C., Tan, W.-C. (2022, June 10). Annotating Columns with Pre-trained Language Models. ArXiv.org. <https://doi.org/10.1145/3514221.3517906>
- [48] Zhou, L., Martínez-Plumed, F., Hernández-Orallo, J., Ferri, C., Schellaert, W. (n.d.). Reject Before You Run: Small Assessors Anticipate Big Language Models. <https://ceur-ws.org/Vol-3169/paper4.pdf>
- [49] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D. (2022, November 21). Language Models (Mostly) Know What They Know. ArXiv.org. <https://doi.org/10.48550/arXiv.2207.05221>
- [50] Wilcoxon, F.: Individual comparisons by ranking methods. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics. Springer Series in Statistics, pp. 196–202. Springer, New York (1992). https://doi.org/10.1007/978-1-4612-4380-9_16
- [51] Team, Q. (2024, March 28). Qwen1.5-MoE: Matching 7B Model Performance with 1/3 Activated Parameters. Qwen. <https://qwenlm.github.io/blog/qwen-moe/>
- [52] Introducing Claude 3.5 Sonnet Anthropic. (n.d.). <https://www.anthropic.com/news/claude-3-5-sonnet>

APPENDIX A

Three Metacognition Scales

In this chapter, we present the results when applying the three metacognition scale to the different datasets: the metacognition-loaded datasets and for the contrast datasets (non-metacognition-loaded), and we will compare the results to when using the single metacognition scale.

A.1 Metacognition-loaded datasets

A.1.1. GPT Family

When applying the three metacognition scale to the metacognition-loaded datasets and using the GPT family for instance outputs, we observe that the AUC values are generally consistent with those obtained using the single metacognition scale, but there are some slight differences.

For the combined benchmarks in Table A.1, the predictability for GPT-3.5 has improved, while for GPT-4, it has slightly decreased. And the same as with the single metacognition scale, GPT-3.5 still achieves a higher AUC value than GPT-4. This suggests that the three metacognition scales may capture certain nuances that are more effectively leveraged by GPT-3.5.

When examining the benchmarks individually in Table A.2, the predictability results remain very similar to those obtained with the single metacognition scale. However, there is a slight improvement in predictability for GPT-4. The KU benchmark continues to exhibit the highest predictability, achieving an AUC of 0.886.

The slight improvements in predictability when using the three metacognition scales suggest that these scales may provide a more nuanced assessment of metacognitive demands. This could be due to the fact that the three scales capture different dimensions of metacognition, offering a more comprehensive evaluation of the cognitive processes involved.

Model	AUC	Brier score
GPT3.5	0.754	0.197
GPT4	0.657	0.230

Table A.1: Assessor performance on the metacognition-loaded datasets evaluated using the three metacognitions scale, with the outputs of the models GPT-3.5 and GPT-4, with results aggregated across all benchmarks.

Benchmarks	Model	AUC	Brier score
EIE	GPT3.5	0.550	0.258
	GPT4	0.522	0.299
KU	GPT3.5	0.552	0.155
	GPT4	0.886	0.086
VfV	GPT3.5	0.518	0.245
	GPT4	0.625	0.239

Table A.2: Assessor performance on the metacognition datasets evaluated using the three metacognition scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.

A.2 Contrast datasets

In this section we will apply the three metacognition scale to the contrast datasets, which are datasets that contain minimal metacognition.

A.2.1. Original HELM models

When applying the three metacognition scale to the original HELM models’ instance outputs, we observe a decrease in predictability for the combined benchmarks. This trend is evident for both the Microsoft TNLGv2_530B model and the aggregated results of all models combined (Table A.3). This decline suggests that the three metacognition scales may not be as effective in capturing the nuances of tasks within the HELM benchmarks, which are designed to have minimal metacognitive content.

When examining the individual benchmarks (Table A.4), the predictability results remain largely consistent with those obtained using the single metacognition scale, there is no significant improvement, indicating that the additional dimensions provided by the three metacognition scales do not offer a substantial advantage for these particular datasets.

The lack of improvement in predictability when using the three metacognition scales may imply that these scales are more suited to tasks with explicit metacognitive demands. Since the HELM benchmarks are designed to assess general language understanding and reasoning without a strong focus on metacognition, the three metacognition scales may not align well with the inherent task characteristics. Consequently, the single metacognition scale appears to be just as effective, for evaluating these non-metacognitive tasks.

Model	AUC	Brier score
All models	0.547	0.258
microsoft/TNLGv2_530B*	0.524	0.249

Table A.3: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the three metacognition scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results aggregated across all benchmarks.

Benchmarks	Model	AUC	Brier score
MMLU Abstract Algebra	All models	0.542	0.255
	microsoft/TNLGv2_530B	0.468	0.279
MMLU Computer Security	All models	0.592	0.257
	microsoft/TNLGv2_530B	0.661	0.232
MMLU College Chemistry	All models	0.603	0.242
	microsoft/TNLGv2_530B	0.618	0.251
MMLU Econometrics	All models	0.554	0.258
	microsoft/TNLGv2_530B	0.442	0.288
MMLU US Foreign Policy	All models	0.616	0.253
	microsoft/TNLGv2_530B	0.478	0.268
BBQ	All models	0.483	0.349
	microsoft/TNLGv2_530B	0.591	0.252
TruthfulQA	All models	0.567	0.265
	microsoft/TNLGv2_530B	0.482	0.263

Table A.4: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the three metacognition scale, with the outputs of the original HELM models and Microsoft’s TNLGv2_530B Model, with results disaggregated by individual benchmarks.

A.2.2. GPT Family

When applying the three metacognition scales to the instance outputs of the GPT family, we observe a general increase in predictability for the combined benchmarks (Table A.5). The assessors achieved higher predictability compared to the single metacognition scale. This indicates that the three metacognition scales capture additional dimensions of metacognitive demand that are relevant even in datasets with minimal metacognitive content.

When examining individual benchmarks (Table A.6), the predictability results varied. For most benchmarks, there was a decrease in predictability compared to the single metacognition scale. However, for the College Chemistry benchmark from the MMLU dataset and the TruthfulQA benchmark, the predictability has improved. This suggests that certain domains within the HELM benchmarks may still benefit from a more detailed metacognitive assessment, even if the overall metacognitive content is low. The improved predictability for the College Chemistry and TruthfulQA benchmarks may indicate that these tasks may involve elements of metacognitive processing that are better captured by the three metacognition scales. For instance, TruthfulQA may involve discerning the veracity of statements, which could engage metacognitive skills.

Model	AUC	Brier score
GPT3.5	0.626	0.242
GPT4	0.686	0.226

Table A.5: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the general difficulty scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.

Benchmarks	Model	AUC	Brier score
MMLU Abstract Algebra	GPT3.5	0.484	0.294
	GPT4	0.471	0.296
MMLU Computer Security	GPT3.5	0.659	0.231
	GPT4	0.453	0.241
MMLU College Chemistry	GPT3.5	0.723	0.214
	GPT4	0.702	0.213
MMLU Econometrics	GPT3.5	0.498	0.293
	GPT4	0.541	0.271
MMLU US Foreign Policy	GPT3.5	0.453	0.246
	GPT4	0.271	0.230
BBQ	GPT3.5	0.514	0.251
	GPT4	0.482	0.260
TruthfulQA	GPT3.5	0.679	0.215
	GPT4	0.521	0.247

Table A.6: Assessor performance on the contrast datasets (non-metacognition-loaded) evaluated using the three metacognition scale, with the outputs of the models GPT-3.5 and GPT-4, with results disaggregated by individual benchmarks.

APPENDIX B

Sustainable Development Goals

Our research aligns with several Sustainable Development Goals (SDGs) set by the United Nations. By advancing our understanding of AI metacognition and developing robust frameworks for evaluating AI performance, this research contributes to the following SDGs:

Sustainable Development Goals	High	Medium	Low	Not applicable
SDG 1. No poverty.				X
SDG 2. Zero hunger.				X
SDG 3. Good health and well-being.				X
SDG 4. Quality education.		X		
SDG 5. Gender equality.				X
SDG 6. Clean water and sanitation.				X
SDG 7. Affordable and clean energy.				X
SDG 8. Decent work and economic growth.				X
SDG 9. Industry, innovation, and infrastructure.	X			
SDG 10. Reduced inequalities.			X	
SDG 11. Sustainable cities and communities.				X
SDG 12. Responsible consumption and production.				X
SDG 13. Climate action.				X
SDG 14. Life below water.				X
SDG 15. Life on land.				X
SDG 16. Peace, justice, and strong institutions.				X
SDG 17. Partnerships for the goals.		X		

1. **SDG 4: Quality Education:** The development of AI systems with enhanced metacognitive abilities can significantly impact the field of education. AI-driven educational tools that can self-evaluate and adapt to the needs of individual learners can provide personalized learning experiences and enhance critical thinking, and thereby improving educational outcomes.

2. **SDG 9: Industry, Innovation, and Infrastructure:** Our research promotes innovation in the field of artificial intelligence, driving the development of more advanced and capable AI systems. By creating predictive frameworks and robust evaluation tools for AI systems, we contribute to the development of resilient infrastructure. These advancements can lead to the creation of smarter, more efficient systems that drive progress in various fields.
3. **SDG 10: Reduced Inequalities:** In the long-term, AI systems with improved metacognitive abilities can help reduce inequalities by providing more inclusive and accessible technologies. For instance, AI-driven educational tools can offer personalized learning experiences to students from diverse backgrounds, helping to bridge the educational gap and ensuring that the benefits of AI are distributed more equitably.
4. **SDG 17: Partnerships for the Goals:** The collaborative nature of this research, conducted in partnership with the OECD AI Skills Reasoning Group, highlights the importance of global partnerships in achieving the SDGs. By working together with international organizations, academic institutions, and industry leaders, we can pool resources, knowledge, and expertise to drive the development of AI technologies that support sustainable development.