ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240770

Language Task Difficulty Prediction Through LLM-Annotated Meta-Features

Yael Moros-Daval[*], **Fernando Martínez-Plumed and José Hernández-Orallo**

UPV - Universitat Politècnica de València

Abstract. Assessing the capabilities of large language models (LLMs) is increasingly challenging due to their generality and uneven task performance. Often, we do not know how much of the success or failure on a particular task is due to the 'loading' of the *language* elements in the task, such as narrative understanding, or some other intrinsic (non-linguistic) components, such as domain-specific common sense or reasoning capabilities. Understanding what tasks are most loaded on language and determine the predictability of LLMs on these tasks is crucial for improving benchmarks, designing better LLMs, and ensuring their safe deployment. We present an innovative methodology that uses LLMs to annotate linguistic meta-features , allowing us to predict task difficulty and understand linguistic loadings more accurately than traditional readability scores. Using GPT-4 for automated annotation, we show strong predictability for a variety of tasks and language models (e.g., MMLU with R^2 from 0.68 to 0.83), but observe limited predictability for other tasks (e.g., LSAT with R^2 of -0.07).

1 Introduction

The progress in large language models (LLMs) [13, 3, 37] is pushing the boundaries of what machines can perform using natural language [11]. These models can be instructed to do a range of unanticipated tasks following the user request, in what is known as human-centred generality [44]. However, their performance varies considerably from task to task [47, 28]. This raises a crucial question: *Is the variation in performance due to the linguistic complexity of the tasks (morphology, syntax, semantics or pragmatics), some other inherent elements (non-linguistic demands, domain-specific knowledge or skills, etc.), or a combination of both?*

Traditional benchmarks, using readability and lexical complexity metrics, have long been used as indicators of human text comprehension [16, 20, 32, 8]. However, they have not been thoroughly evaluated against the diverse and demanding requirements of LLMs [9]. It is crucial to unravel the linguistic barriers that LLMs may face, as this knowledge is as important as assessing their reasoning capabilities [24]. Consequently, this research is driven by two goals: to provide a critical assessment of conventional text difficulty metrics in the context of LLMs, and to conceptualise an innovative approach to meta-feature annotation that uses the LLMs themselves to generate an extensive annotated text repository.

Challenging the conventional task-by-task approach in evaluating AI systems, the proposed paradigm adopts an instance-level analytical framework [6]. It advocates the detailed annotation of task instances with meta-features that reflect the cognitive demands of each task, providing a granular view of AI performance through capability-based evaluation [23]. Assigning difficulty to instances has a wide range of applications, such as improving benchmarks, predicting performance for new instances and tasks, understanding the loading and variance of benchmarks, etc. To validate these meta-features, we perform an empirical analysis using them as predictors of the difficulty LLMs find on the task. It builds incrementally on the foundation laid by previous readability metrics, addressing their limitations and proposing an advanced rubric that captures the multifaceted nature of linguistic complexity. In doing so, the paper evaluates the hypothesis that well-defined linguistic meta-features can effectively anticipate where LLMs may struggle linguistically, and not because of other elements of the benchmark.

Our key contributions include:

1. A comprehensive **review of classical lexical complexity and readability metrics**, evaluating their utility as reliable predictors of task difficulty for LLMs.
2. A complete **set of linguistic meta-features** that capture the inherent demands that LLMs encounter in natural language.
3. A validated **automated annotation methodology** using LLMs for efficient linguistic meta-feature labelling across benchmarks such as BIG-bench [47] and HELM [28].
4. A demonstration of the efficacy of linguistic meta-features in **predicting NLP task difficulty** for tasks with high loadings of linguistic elements.

After the introduction, the paper reviews the relevant literature, details the annotation framework, and describes the experimental setup and results. Conclusions close the paper.

2 Background

NLP Tasks Evaluation. Computational linguistic and Natural Language Processing (NLP) have focused on the computational analysis and representation of human language and the construction of systems that solve particular language tasks [29, 10]. Different morphological, lexical, syntactic and semantic aspects were studied and evaluated, and there was a clear division between tasks, even at the abstract level of separating understanding and generation components. However, language models have blurred all these distinctions. Still, inheriting from a decades-long tradition of NLP benchmarks, modern LLM benchmarks such as BigBench [47] or HELM [28] include several NLP tasks, facilitating the comparison of LLMs between them and with some other NLP systems [28]. However, even

[*] Corresponding Author. Email: ymordav@inf.upv.es.

benchmarks that focus on language understanding can contain other cognitive aspects such as reasoning or knowledge, as tasks commonly extend beyond 'pure language'.

Lexical and readability metrics. Dating back to the early 20th century, readability assessments (for humans) have traditionally focused on vocabulary and syntax, excluding aspects such as semantics and discourse. Despite their simplistic approach, exemplified by metrics such as the Flesch Reading Ease Score [16], these early formulas have remained competitive against newer, more complex models [17]. With advances in NLP and machine learning, more integrated tools such as Coh-Metrix have emerged, combining language features with cognitive and discourse analysis to improve readability predictions [19]. The comprehensive framework developed by Graesser et al. serves as an important source of inspiration for our meta-feature identification process, although it does not fully address the non-propositional aspects of language, such as modality and negation. In general, it has been well recognised [33] there is a need for rich and adaptable metrics to understand and annotate linguistic complexity, and libraries are incrementally updated with new metrics every year.

Large Language Models (LLMs) evaluation. LLMs such as GPT-4 [37] exemplify the latest advances in language model capabilities, driven by deep learning transformer architectures [50]. The performance of these models seem to scale up with dataset size, learning time and, most especially, number of parameters, but this increase is very uneven across tasks [46, 25, 1, 7]. Also, the traditional methodologies and metrics for task-by-task evaluation are increasingly inadequate for these general-purpose systems, as they often overlook the complexity required in various tasks, which is determined by different mixtures of difficulty, demands, or specific knowledge skills needed for addressing the benchmarks. This highlights the need for more nuanced, instance-level analysis to truly assess their diverse capabilities [5].

Automated annotation using LLMs. With the increasing demand for annotated text data in machine learning, the use of LLMs as automated annotators is becoming a viable alternative to manually-intensive annotation processes [22, 43, 48]. Few-shot learning allows LLMs to quickly grasp annotation criteria from examples or rubrics, generating large datasets efficiently and consistently. Their adaptability means they can be finetuned (via new training examples or carefully designed prompts) to meet project-specific needs, saving time and resources in large-scale annotation efforts [3]. Still, to ensure the highest level of accuracy, researchers usually complement the LLM-generated annotations with human verification [38].

3 Automated Annotation of Demands

The core of our proposed framework is to annotate each task instance with meta-features representing its *instance demands*.

As we annotate the datasets, these meta-features allow us to make more sophisticated connections between tasks and abilities. Similar to Item Response Theory (IRT) [21], we can consider all meta-features mapping to a single proxy for task difficulty and derive the 'ability' as a latent factor [14, 30, 31]. However, instead of inferring difficulty from performance data, our approach predicts task difficulty directly from the linguistic meta-features. This predictive modelling allows us to anticipate where LLMs may encounter challenges, based on a nuanced understanding of task-specific linguistic demands. In this regard, Figure 1 contrasts the predictive value of linguistic meta-features (composition and space) with that of more traditional non-predictive readability metrics (SMOG and FORCAST). The linguistic meta-features seem to have more potential as predictors of task difficulty.

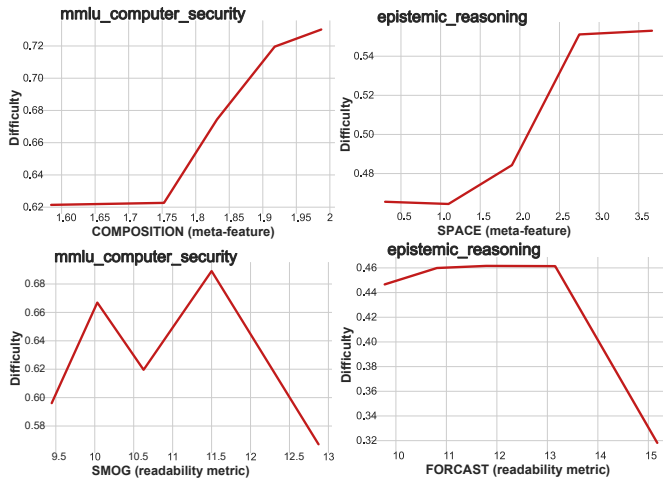

Figure 1: Characteristic curves showing the predicted average task difficulty in relation to the (binned) demands of linguistic meta-features (top: composition and space) compared to traditional readability metrics (bottom: SMOG and FORCAST) on *MMLU Computer Security* and *Epistemic Reasoning* tasks.

But how do we know the demand levels for each instance? This is the role of annotations. Consider, for example, a question-answering dataset with different questions, each requiring different levels of reasoning and vocabulary. The specific abilities required by each question can be annotated accordingly:

Annotation example

Q_1 "In a game of chess, when a player's knight is under attack, the player has several possible moves to consider. What are the different options available to the player?"

- Level of: **Reasoning** (2); **Vocabulary** (0.2)

Q_2 "What is the exact definition of the word sesquipedalian?"

- Level of: **Reasoning** (0); **Vocabulary** (0.7)

Given this new setting for the estimation of difficulties, there are two ways in which we can have annotated benchmarks: one involves meticulously creating benchmarks from scratch based on cognitive concepts, while the other involves retrospectively annotating existing benchmarks. There are notable examples in the literature where researchers examine (or argue in favour of) the elicitation of the capabilities that are required for a given task, and develop targeted tests to assess models across these different skills [42, 45, 18, 12, 35]. Conversely, instance annotation can be used when this process has not been possible or to complement it. Traditionally, this method relies on human experts, who are prone to human error and bias. The use of LLMs for automated annotation offers a more efficient and scalable alternative.

3.1 Meta-Feature Definition

The backbone of our method is the annotation of linguistic tasks with meta-features that reflect task requirements at the instance level.

Traditional readability metrics often overlook the complex linguistic and cognitive demands of these tasks. The proposed meta-features (see Table 1) are not only linguistic, involving syntax and semantics such as 'negation' and 'compositionality', but some also encapsulate non-linguistic aspects such as 'reasoning', thus extending the evaluative scope beyond language alone, provided they are generic, i.e., domain-independent.

The selection of meta-features in our framework is influenced by their established relevance in contemporary linguistic and cognitive research. Each feature has been chosen for its proven impact on language comprehension, processing or production as observed in empirical studies, ensuring that they can be used for evuating the demands of any task expressed with language. There are meta-features related to linguistic components. For example, 'negation' has been extensively studied in psycholinguistics due to its impact on the complexity of sentence processing and is known to increase cognitive load [51, 26]. 'Modality' is similarly important as it affects the inferential processes involved in language comprehension, a subject of study in formal semantics and pragmatics [27]. Complementarily, some meta-features focus on cognitive dimensions, that are not captured by traditional metrics. Theory of mind, another meta-feature, is critical for processing narrative and understanding the intentions behind speech acts, a key issue in developmental psychology and narrative theory [40, 4]. 'Reasoning', while arguably a broader cognitive skill, is implicated in language when understanding arguments and logical structures within texts [36]. 'Compositionality' is directly related to semantic theory, showing how the meaning of a complex expression is determined by its parts and their syntactic combination [39]. These meta-features are also consistent with efforts in the evaluation of language models, where researchers are increasingly recognising the need for more granular, feature-specific performance evaluations [15].

The concept of meta-features in this paper not only identifies features that affect performance on linguistic tasks, but also establishes quantifiable scales for each of them. While the scaling of some features, such as the proportions of certain word types within a text, is clear, the assessment of non-propositional elements, such as negation, adds subjective complexity to the scaling. This can range from a binary measure of absence or presence to more granular scales that take into account frequency within a text, as illustrated by the double use of negation in sentences such as "*Neither Paula nor Maria is going to their respective entertainment events*".

The challenge lies in assigning clear numerical values to each meta-feature without arbitrary decisions. The granularity at which the meta-feature is calculated, i.e., whether it is the whole text, individual paragraphs or sentences, can also affect the result, as can language dependencies. The list of meta-features selected for this study (Table 1) includes both clear procedural rules for calculation, such as for 'negation' or 'compositionality', and more subjective judgements based on example 'anchors'—exemplary cases that provide reference points for classification within categories (see Appendix C in [34] for examples). Where rules are not clear, such as determining the level of uncertainty in a sentence, manual annotation may be the only viable option. This inherently subjective process lacks concrete ground truth, making these levels a convention rather than absolute measures.

Is this meta-feature approach better than the use of traditional readability metrics? Our study seeks to juxtapose the two and examine their effectiveness across AI benchmarks of varying complexity and skill requirements.

Table 1: Description of linguistic meta-features. See Appendix C in [34] for further details on the scale and examples.

Meta-feature	Definition (Scale)
Uncertainty	Refers to epistemic situations involving imperfect or unknown information. (**0...10**)
Negation	Refers to a denial, contradiction, or negative statement. (**0...**)
Time	A temporal expression in a text is a sequence of tokens (words, numbers and characters) that denote time, duration, or frequency. (**0...**)
Space	A spatial expression in a text is a sequence of tokens (words, numbers and characters) relating to the position, area, and size of things. (**0...**)
Vocabulary	The vocabulary level is measured by a normalized metric of the log frequency of words. (**0...1**)
Modality	Refers to a classification of propositions on the basis of whether they claim necessity, possibility, or impossibility. (**0...**)
Theory of Mind	In psychology, theory of mind refers to the capacity to understand other people by ascribing mental states to them. (**0...**)
Reasoning	Is the process of forming conclusions, judgments, or inferences from facts or premises. (**0...**)
Compositionality	In semantics, the principle of compositionality states that the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them. (**1...levels**)
Anaphora	Is the use of a linguistic unit, such as a pronoun, to refer to the same person or object as another unit, usually a noun. (**0...**)
Noise	Is the number of typos per character with respect to the original text with no typos. (**0...typos**)

3.2 *Prompt design*

The annotation methodology uses LLMs and few-shot learning to annotate numerous examples using predefined linguistic meta-features. Prompts are carefully crafted to contextualise meta-features and illustrate their scales, followed by sentences for model annotation. Models are suggested to take the role of expert in linguistics within these prompts, using scales and example-based definitions for meta-features to guide the annotation task (see Appendix C in [34]). The distinct attributes of each meta-feature, such as 'uncertainty', 'negation' and 'modality', are used to inform the annotations, with the full annotation template detailed in the study (see the template in the following box).

Prompt template

```
You are a helpful expert on linguistics. You must
help me annotate the level of {META-FEATURE}
of a given sentence/s. Note that {META-FEATURE
DEFINITION}. I will first give you a few examples
to illustrate it (as few-shot learning). Then you
will have to determine the level of {META-FEATURE}
for a new sentence/s on a scale from {META-FEATURE
SCALE}.
{META-FEATURE EXAMPLES}
Sentence: {INSTANCE}
Level of {META-FEATURE}:
```

While this approach is advantageous for its automation potential, it is not without its drawbacks. High token costs can be incurred when

processing millions of examples, and there is a need for manual intervention in cases where model responses fall outside the established level range, requiring post-processing. The first step is to create a pool of annotated examples which, despite the automation goals, still require some human input and review.

4 Experimental setting

4.1 Data Repositories.

We have curated a representative collection of datasets from two prominent repositories: BIG-bench and HELM. These datasets provide extensive coverage across various domains and models, enabling us to analyse the nuanced performance differences of AI on a wide array of tasks. The Beyond the Imitation Game benchmark (BIG-bench) serves as a comprehensive repository for over 200 distinct tasks. From this pool, we selected tasks that presented signs of variable difficulty, identified by metrics such as mean accuracy, variability in question length, and the combined length of the question and its possible answers. We aim to ensure that our sample represents a range of textual complexity and presents different levels of challenge to models. See Table 2 for more information. Our study analyses data from two different families of models: BIG-G T=0 [49] and BIG-G sparse [52], which include models ranging from 2M to 128B parameters, with further details in [34] (Appendix A). The Holistic Evaluation of Language Models (HELM) [28] benchmark emphasises the transparency of language model evaluations. It stands out for its uniform execution of various tasks across major AI models under identical conditions, offering an array of instance-level data. HELM contains more than forty-two scenarios, encompassing domains such as legal reasoning and commonsense question-answering. We selected eight multiple-choice tasks, each described in Table 2, with a comprehensive list of evaluated LLMs with parameter counts between 350M and 540B (details provided in Appendix A in [34]). Ultimately, we capture a representative range of NLP benchmarks. By including tasks from different domains (i.e., narrative comprehension, legal reasoning, common-sense, question-answering, etc.) we sought to evaluate the predictive effectiveness of linguistic meta-features and readability metrics in different contexts, while keeping the experimental setting manageable given the instance-level analysis framework adopted (with approximately 25K instances and over 933K annotations in total).

Difficulty prediction. In our experimental setting, the difficulty of an AI benchmark instance is simply defined as

$$\text{diff} = 1 - \frac{\sum_{i=1}^{N} \text{correct}_i}{N}$$

where correct_i is the result (correct or incorrect) of model i, and N is the total number of models. A difficulty score of zero represents an easy task for the models, while a score of one means that no model has successfully solved the instance, indicating a high level of difficulty. In order to predict this difficulty from metrics and meta-features, we use a Random Forest [2] regressor. The regressor is used with standard parameters from *Scikit-learn*[1], with a 75/25 train-test split. We evaluate the performance of the regressor using the coefficient of determination (R^2) and the RMSE metric. Each task and setting is modelled independently to understand specific task behaviours and to analyse how the demands (meta-features) influence on the prediction of task difficulty.

Readability metrics. One part of the study is to evaluate the ability of existing human readability and lexical diversity metrics to predict the performance of language models. To compute these metrics for each instance, we use the QUANTEDA[2] R package. The text considered for the calculation is the result of concatenating the input question and the correct answer. From the available set of metrics in the literature (see Appendix B in [34]), we selectively identified metrics with minimal collinearity and significant presence in the literature. Preliminary correlation analyses guided our choice, avoiding redundant measures. The correlation matrix for *MMLU Computer Security* in the [34] (Figure 1) demonstrates our selection rationale. Our chosen measures include lexical diversity metrics such as Type-Token Ratio (*TTR*) and Yule's *K*, which assess vocabulary variation. TTR indicates lexical range by the proportion of unique words to total words in a text, while a lower Yule's K indicates less word repetition and greater diversity. On the readability front, we chose metrics such as *Flesch*, *Scrabble*, *FOG*, *SMOG.C* and *FORCAST*. These assess how easy it is to understand text, with the Flesch score measuring ease based on sentence length and syllables, and the Fog and SMOG indices relating readability to education level and sentence complexity. The FORCAST metric is tailored to the readability of technical documents and focuses on the frequency of single-syllable words.

Meta-feature preparation and annotation. In preparing meta-features for our experimental setting, both HELM and BIG-bench question-answer inputs were systematically broken down into individual sentences using NLTK's sentence tokenizer[3]. This approach decomposes complex inputs into simpler components such as distinct sentences and answer choices. To annotate the meta-feature values, we used GPT-4 [37] , with the "temperature" parameter set to zero to ensure that the results were deterministic.

Postprocessing. In the post-processing phase, we clip the output of the language model to ensure that all annotated meta-feature values adhere to the predefined scales, setting any out-of-range values to their respective minimum or maximum scale limits. After ensuring that all individual sentence values are appropriately scaled, we then determine the collective meta-feature value for complete instances, typically paragraphs consisting of multiple sentences. We compute the average level across sentences, which provides an overall measure of meta-feature prevalence in a paragraph. Detailed figures for meta-feature ranges for specific tasks are included in Appendix D in [34].

5 Results

A bird-eye's view of the results on two metrics, R^2 (Table 3) and RMSE (Figure 6) shows that the difficulty predictability results are similar for many benchmarks, but with major gains in favour of the meta-features approach for *Epistemic Reasoning*, *Formal Fallacies Syllogism Negation* and *Legal Support*. We need a more thorough evaluation to assess how well linguistic meta-features and readability metrics predict the difficulty of the selected tasks, by analysing R^2 and RMSE separately[4].

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[2] https://quanteda.io/

[3] https://www.nltk.org/index.html

[4] Following the Science paper's guidelines for AI evaluation reporting [6], all code, data and instance-level results are available at https://github.com/yaelmd/llm-metafeatures.

Table 2: Description of selected BIG-bench and HELM tasks. Each of the tasks contains more than one thousand instances.

Bench	Task	Description	Domain
BigBench	Abstract Narrative Understanding	Given a narrative, choose the most related proverb	Analogical Narrative understanding Social reasoning
BigBench	Formal Fallacies Syllogisms Negation	Distinguish deductively valid arguments from formal fallacies	Fallacy Logical reasoning Negation
BigBench	Epistemic Reasoning	Determine whether one sentence entails the next	Common sense Logical reasoning Social reasoning Theory of mind
HELM	Massive Multitask Language Understanding (MMLU)	Knowledge-intensive question answering across 4 domains: Computer Security, US Foreign Policy, Econometrics and College Chemistry	Knowledge-intensive QA
HELM	OpenbookQA	Commonsense-intensive open book question answering	Knowledge-intensive QA
HELM	Legal Support	Fine-grained legal reasoning through reverse entailment	Legal Realistic Reasoning
HELM	LSAT	Measure analytical reasoning on the Law School Admission Test	Logical Realistic Reasoning
HELM	Bias Benchmark for Question Answering (BBQ)	Social bias in question answering in ambiguous and unambiguous context	Bias
HELM	HellaSwag	Commonsense reasoning in question answering	Knowledge-intensive QA
HELM	TruthfulQA	Model truthfulness and commonsense knowledge in question answering	Knowledge-intensive QA

Table 3: R^2 obtained in the test split when predicting difficulty with linguistic meta-features and lexical and readability metrics.

Task	Meta-features	Readabilitiy
Abstract Narrative Understanding	0.06	-0.01
BBQ	0.62	0.5
Epistemic Reasoning	0.9	-0.03
Formal Fallacies Syllogisms Negation	0.6	-0.15
Hellaswag	0.02	-0.03
Legal Support	0.3	0.05
LSAT	-0.07	-0.07
MMLU College Chemistry	0.77	0.74
MMLU Computer Security	0.83	0.85
MMLU Econometrics	0.68	0.7
MMLU US Foreign Policy	0.8	0.83
OpenbookQA	-0.04	0.01
TruthfulQA	0.59	0.56

5.1 $\mathbf{R^2}$ *analysis*

We can categorise tasks into three groups based on their predictability, as indicated by the range of R^2 values: (1) those highly predictable by both approaches, (2) those better predicted by linguistic meta-features alone, and (3) tasks for which neither approach predicts well.

The first group, which is composed of tasks with high $R^2 \geq 0.5$ for both approaches, includes *BBQ*, *MMLU College Chemistry*, *MMLU Computer Security*, *MMLU Econometrics*, *MMLU US Foreign Policy* and *TruthfulQA*. In Figure 2, we can see that, e.g., for *MMLU US Foreign Policy*, actual difficulty and predicted one are pretty similar. This result makes it clear that these tasks have a linguistic base, and because of that, we can predict their difficulty both with linguistic meta-features and traditional metrics. The question of which of the two approaches is better will be addressed below.

Then there are those datasets whose difficulty is not well predicted by traditional metrics, *Epistemic Reasoning* and *Formal Fallacies Syllogisms Negation*. We can take a look at Figure 3 to see the difference between meta-features and readability metrics predictions in *Epistemic Reasoning*. Maybe for these tasks, linguistic meta-features are able to capture some relevant characteristics that traditional metrics do not.

Finally, there are some tasks where neither readability metrics nor meta-features properly predict difficulty. These are *Abstract Narrative Understanding*, *Hellaswag*, *LSAT*, *Legal Support* and *OpenbookQA*. In Figure 4 we observe the results of *LSAT*, the model predictions do not fit with the actual ones.

The limited number of models evaluated —only six for tasks such as *Hellaswag* and *OpenbookQA*— may contribute to their poor predictability of difficulty scores, in contrast to other tasks which have results from at least 32 models. This disparity is illustrated in Figure 5, where the sparsity of data points results in apparent vertical lines on the difficulty graph. The lack of other discernible patterns suggests that these tasks may not have a strong linguistic basis, having mixed requirements of more advanced cognitive abilities that are not fully captured by either linguistic meta-features or traditional readability metrics, making it hard to predict their linguistic difficulty.

less about linguistic complexity than about the reasoning or domain-specific knowledge required

5.2 *RMSE analysis*

The RMSE results, detailed in Figure 6, reveal distinct patterns across tasks, allowing us to categorise them according to which pre-

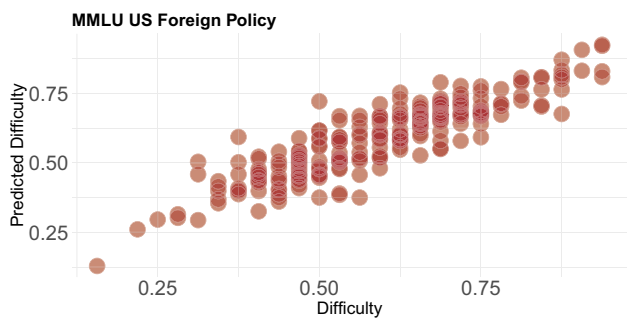

Figure 2: Predicted Difficulty vs. Actual Difficulty for *MMLU US Foreign Policy* using linguistic meta-features.

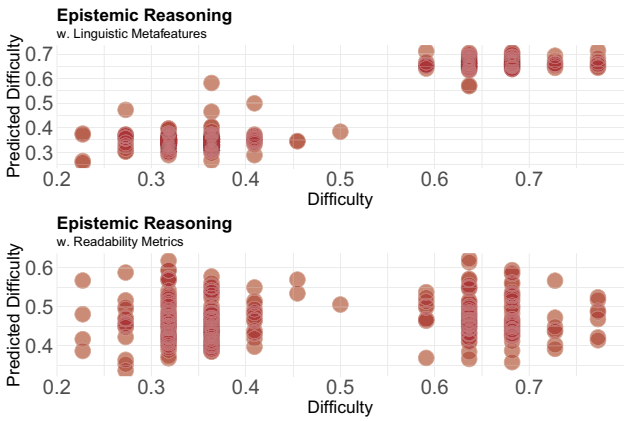

Figure 3: Predicted Difficulty vs. Actual Difficulty for *Epistemic Reasoning* using linguistic meta-features (top), lexical and readability metrics (bottom).

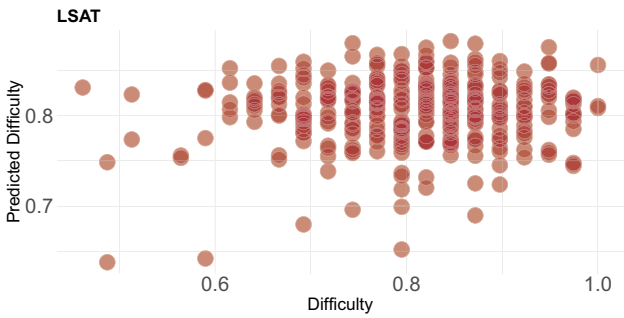

Figure 4: Predicted Difficulty vs. Actual Difficulty for *LSAT* using linguistic meta-features.

dictive approach yields lower RMSE values: **(Cat. 1)** tasks for which **linguistic meta-features yield lower RMSE**, indicating better predictability; **(Cat. 2)** tasks where **lexical diversity and readability metrics are more predictive**; and **(Cat. 3)** tasks with **comparable predictiveness** from both linguistic meta-features and readability metrics. This classification implies that while linguistic meta-features excel at predicting difficulty for many tasks, their superiority is not universal, highlighting the need to also consider task-specific features.

Starting with those **tasks with lower RMSE using linguistic meta-features (Cat. 1)**, in the tasks analysed, linguistic meta-features proved more effective than traditional metrics for predicting difficulty, with significantly lower RMSE values. Most of the tasks in this group are classified as purely reasoning tasks as we can see in Table 2 (e.g., *Abstract Narrative Understanding* and *Epistemic Reasoning*).

Epistemic Reasoning example

Identify the **relation** between the following premises and hypotheses, choosing from the options 'entailment' or 'non-entailment'.
Premise: Charles remembers that Amelia believes that a black scientist looks through a scope examining a biological specimen's blood cells.
Hypothesis: Amelia believes that a black scientist looks through a scope examining a biological specimen's blood cells.

Even knowledge-intensive tasks such as *Hellaswag* benefited from the improved accuracy provided by meta-features, highlighting their

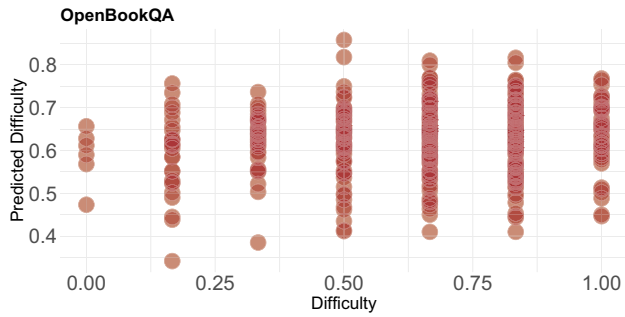

Figure 5: Predicted Difficulty vs. Actual Difficulty for *OpenbookQA* using linguistic meta-features.

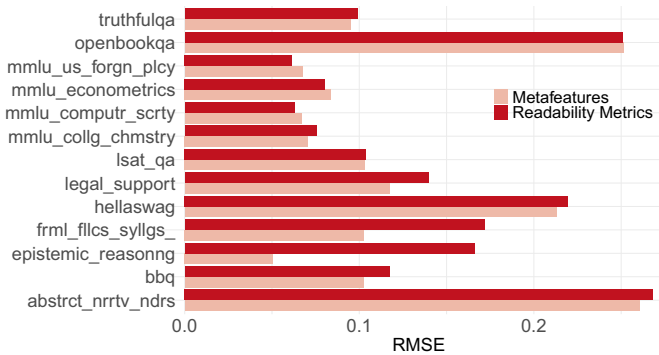

Figure 6: RMSE values obtained from predicting text difficulty for each task (names abbreviated) using linguistic meta-features and lexical diversity and readability metrics.

depth in capturing commonsense knowledge. It assesses a model's ability to understand and reason about real-world scenarios, requiring an appreciation of context, purpose and the implicit meanings within language use.

This analysis suggests that linguistic meta-features provide a sharper tool for gauging the difficulty that AI models might face in language tasks, capturing complexities that go beyond lexical diversity and readability.

The second group of **tasks with lower RMSE when using traditional lexical diversity and readability metrics (Cat. 2)** include *MMLU Computer Security*, *MMLU US Foreign Policy* and *MMLU Econometrics*. These tasks generally involve simpler language processing and rely more heavily on vocabulary and comprehension skills, so traditional metrics such as Flesch Reading Ease or FORCAST.RGL are more appropriate for estimating difficulty.

All of these tasks in the HELM benchmark fall into the category of knowledge question-answering and hence should require domain-specific theoretical knowledge rather than complex reasoning. An example from the *MMLU Computer Security* item illustrates this point:

MMLU Computer Security Question

Exploitation of the Heartbleed bug permits
A. overwriting cryptographic keys in memory
B. a kind of code injection
C. a read outside bounds of a buffer
D. a format string attack
Answer: **C**

However, answering such questions requires detailed conceptual understanding, but not higher cognitive skills or very sophisticated

Table 4: Average ranking of each meta-feature in feature importance for each category.

Meta-feature	Cat. 1	Cat. 2	Cat. 3
Uncertainty	2.12	2.67	3.00
Negation	7.50	3.33	10.50
Time	7.25	8.67	7.5
Space	7.00	6.67	5.00
Vocabulary	1.75	1.00	1.00
Modality	6.88	6.67	6.00
Theory of Mind	7.00	9.00	10.50
Reasoning	6.38	7.67	8.50
Compositionality	6.63	6.33	3.00
Anaphora	6.38	10.00	8.00
Noise	7.12	4.00	3.00

specific knowledge (or at least this is not making the difference between success and fail). This is why they are named 'language understanding' tasks, and most of the MMLU items share this characteristic. The only exception is College Chemistry item, which requires more applied knowledge, and the loading of language understanding is hence lower. Different types of knowledge requirements between domains —conceptual versus practical— may contribute to the variation in the predictability of task difficulty as reflected in the RMSE.

Finally, and regarding the **tasks with similar RMSE for both feature sets (Cat. 3)**, in the realm of AI task difficulty prediction, *OpenbookQA* and *LSAT QA* stand out as tasks where both linguistic meta-features and traditional lexical/readability metrics yield similar RMSE values. This suggests that the complexity of these tasks stems from a blend of both basic linguistic processing and more advanced cognitive abilities.

OpenbookQA question

A fire started in a forest but it wasn't started by people. What could have been the cause?
A. a careless bird
B. a smoking bear
C. electricity
D. a campfire
Answer: **C**

The example from *OpenbookQA* demonstrates how common sense and a moderate level of reasoning are required to deduce the correct answer—a complexity that neither linguistic meta-features nor readability metrics alone can wholly capture. The notable importance of compositionality and noise for these tasks further reflects their unique demands, combining language structure intricacies with potential ambiguities or irregularities in text.

5.3 Feature importance analysis

Finally, it is relevant to note down what the majority of the tasks have in common. Analysis of feature importance across tasks revealed common influential factors in predicting difficulty. Vocabulary and uncertainty emerged as the two most critical meta-features, carrying the highest average importance ranks of 1.46 and 2.38, respectively. In the realm of traditional metrics, *Scrabble* (reflecting word complexity), *TTR* (gauging lexical variation), and Yule's *K* (measuring vocabulary richness) were identified as key indicators, each demonstrating significance across all tasks with respective average ranks of 2.31, 2.54, and 2.77 (as shown in Table 5). These patterns suggest a strong lexical influence on task difficulty, highlighting the importance of diverse and certain vocabulary in determining how challenging a task is deemed.

Table 5: Average Ranking Position in feature importance from the four best positioned meta-features (left) and readability/lexical metrics (rigth).

Meta-feature	Average Rank	Metric	Average Rank
Vocabulary	1.46	Scrabble	2.31
Uncertainty	2.38	TTR	2.54
Noise	5.77	K	2.77
Compositionality	6	FORCAST.RGL	3.54

6 Conclusions

The ultimate aim of this paper was to showcase the usefulness of a new meta-feature approach to assessing the difficulty of natural language tasks. Our results confirm that the defined meta-features can predict task difficulty and provide deeper insights into the demands of individual instances. However, the relative performance of meta-features and traditional readability metrics varies across tasks. Still, we were able to identify unique linguistic and cognitive demands for each task type, highlighting the nuanced nature of task difficulty.

We have also shown that task annotation using language models is effective and feasible, as evidenced by the successful annotation and analysis of 13 tasks involving thousands of instances. While the accuracy of each individual automated annotation cannot be absolutely guaranteed, it is typically on par with human annotation accuracy [41].

Future research could first diversify the results by exploring a wider range of tasks and language models, while also investigating the effects of integrating additional linguistic complexity and cognitive demands. For instance, it would be very insightful to see if there are domain-specific demands that give a push of predictability in those tasks where language is not sufficient to predict the outcome. In the end, this methodology allows us to digest the sources of difficulty, and ultimately, the sources of error. In addition, the underlying concept of annotating tasks with meta-features that reflect their inherent demands can be extended to other domains. The methodology can be easily applied to accommodate for more metafeatures when dealing with datasets that require specific knowledge or new skills. Even, this methodology can be applied beyond LLMs, for other multimodalities. For instance, given images representing a task, we could use rubrics to ask multimodal models to annotate the demands of the tasks, and then build predictive models out of them.

Acknoledgements

We acknowledge support from: Grant for Master Studies funded by ValgrAI – Valencian Graduate School and Research Network for Artificial Intelligence and Generalitat Valenciana; FISCALTICS (I+D+i PID2022-140110OA-I00), granted by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; CIPROM/2022/6 (FASSLOW) and IDIFEDER/2021/05 (CLUSTERIA) funded by Generalitat Valenciana; the EC H2020-EU grant agreement No. 952215 (TAILOR); US DARPA HR00112120007 (RECoG-AI) and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and ERDF.

References

[1] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. Explaining Neural Scaling Laws, 2021.

[2] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] J. S. Bruner. *Actual minds, possible worlds*. Harvard university press, 2009.

[5] R. Burnell, J. Burden, D. Rutar, K. Voudouris, L. Cheke, and J. Hernández-Orallo. Not a number: Identifying instance features for capability-oriented evaluation. In *IJCAI*, pages 2827–2835, 2022.

[6] R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela, M. Shanahan, E. M. Voorhees, A. G. Cohn, J. Z. Leibo, and J. Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023. doi: 10.1126/science.adf6369. URL https://www.science.org/doi/abs/10.1126/science.adf6369.

[7] E. Caballero, K. Gupta, I. Rish, and D. Krueger. Broken Neural Scaling Laws. In *The Eleventh International Conference on Learning Representations*, 2022.

[8] J. S. Caylor and T. G. Sticht. Development of a simple readability index for job reading material. 1973.

[9] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[10] K. R. Chowdhary. *Natural Language Processing*, pages 603–649. Springer India, New Delhi, 2020. ISBN 978-81-322-3972-7. doi: 10.1007/978-81-322-3972-7_19. URL https://doi.org/10.1007/978-81-322-3972-7_19.

[11] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*, 2022. doi: 10.48550/arXiv.2210.11416.

[12] I. Dasgupta, D. Guo, A. Stuhlmüller, S. J. Gershman, and N. D. Goodman. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*, 2018.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] S. E. Embretson and S. P. Reise. *Item response theory*. Psychology Press, 2013.

[15] A. Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[16] R. Flesch. Marks of readable style; a study in adult education. *Teachers College Contributions to Education*, 1943.

[17] T. François and E. Miltsakaki. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57, 2012.

[18] M. Gardner, J. Berant, H. Hajishirzi, A. Talmor, and S. Min. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*, 2019.

[19] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234, 2011.

[20] R. Gunning. The technique of clear writing. mcgraw-hill. *New York*, 1952.

[21] R. K. Hambleton and H. Swaminathan. *Item response theory: Principles and applications*. Springer Science & Business Media, 2013.

[22] X. He, Z. Lin, Y. Gong, A.-L. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, and W. Chen. AnnoLLM: Making large language models to be better crowdsourced annotators, 2023.

[23] J. Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48:397–447, 2017.

[24] J. Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.

[25] M. Hutter. Learning Curve Theory, 2021.

[26] B. Kaup, J. Lüdtke, and R. A. Zwaan. Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7):1033–1050, 2006.

[27] A. Kratzer, H.-J. Eikmeyer, and H. Rieser. The notional category of modality. *Formal semantics: The essential readings*, pages 289–323, 1981.

[28] P. Liang, R. Bommasani, T. Lee, et al. Holistic evaluation of language models, 2022.

[29] C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[30] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo. Making sense of item response theory in machine learning. In *ECAI 2016*, pages 1140–1148. IOS Press, 2016.

[31] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271:18–42, 2019.

[32] G. H. Mc Laughlin. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[33] R. Morante and C. Sporleder. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260, 2012.

[34] Y. Moros-Daval, F. Martínez-Plumed, and J. Hernández-Orallo. Supplementary material: Predicting language task difficulty using LLM-annotated meta-features. 2024. URL https://github.com/yaelmd/llm-metafeatures.

[35] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*, 2018.

[36] M. Oaksford and N. Chater. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4):608, 1994.

[37] OpenAI. GPT-4 technical report, 2023.

[38] N. Pangakis, S. Wolken, and N. Fasching. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*, 2023.

[39] B. Partee et al. Compositionality. *Varieties of formal semantics*, 3: 281–311, 1984.

[40] D. Premack. Does the chimpanzee have a theory of mind? *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*, pages 160–179, 1988.

[41] J. Pustejovsky and A. Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.", 2012.

[42] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

[43] J. Savelka and K. D. Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6, 2023.

[44] W. Schellaert, F. Martínez-Plumed, K. Vold, J. Burden, P. A. Casares, B. S. Loe, R. Reichart, A. Korhonen, J. Hernández-Orallo, et al. Your prompt is my command: On assessing the human-centred generality of multimodal models. *Journal of Artificial Intelligence Research*, 77:377–394, 2023.

[45] D. Schlangen. Targeting the benchmark: On methodology in current natural language processing research. *arXiv preprint arXiv:2007.04792*, 2020.

[46] U. Sharma and J. Kaplan. A Neural Scaling Law from the Dimension of the Data Manifold, 2020.

[47] A. Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

[48] Y. Suhara, J. Li, Y. Li, D. Zhang, Ç. Demiralp, C. Chen, and W.-C. Tan. Annotating columns with pre-trained language models. In *Proceedings of the 2022 International Conference on Management of Data*, pages 1493–1503, 2022.

[49] R. Thoppilan et al. LaMDA: Language Models for Dialog Applications. *arXiv preprint arXiv:2201.08239*, 2022. doi: 10.48550/arXiv.2201.08239.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] P. C. Wason. The contexts of plausible denial. *Journal of verbal learning and verbal behavior*, 4(1):7–11, 1965.

[52] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, and W. Fedus. ST-MoE: Designing Stable and Transferable Sparse Expert Models. *arXiv preprint arXiv:2202.08906*, 2022. doi: 10.48550/arXiv.2202.08906.