Supplementary Material: Predicting Language Task Difficulty Using LLM-Annotated Meta-Features

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

UPV - Universitat Politècnica de València

This supplementary material serves as technical appendix for the paper *Language Task Difficulty Prediction through LLM-Annotated Meta-Features* [5] published in the European Conference on Artificial Intelligence (ECAI 2024).

A Description of LLMs in the BIG-bench and HELM benchmarks

The BIG-bench benchmark uses a wide range of Large Language Models (LLMs) of varying complexity. These models range from the smaller 2 million parameter tier to expansive architectures with up to 128 billion parameters.The performance and linguistic capabilities of these LLMs are examined across a wide range of tasks, testing their ability to navigate complex language scenarios with nuance and depth. Detailed characteristics of the BIG-G models from the benchmark are catalogued in the table 1, which lists the models according to their size, reflecting the computational heft and potential sophistication in processing natural language tasks.

Table 1: BIG-G models addressing BigBench [6]

Model	#Parameters
BIG-G sparse	2M
BIG-G sparse	16M
BIG-G sparse	53M
BIG-G sparse	125M
BIG-G sparse	244M
BIG-G sparse	422M
BIG-G sparse	1B
BIG-G sparse	2B
BIG-G sparse	4B
BIG-G sparse	8B
BIG-G dense	2M
BIG-G dense	16M
BIG-G dense	53M
BIG-G dense	125M
BIG-G dense	244M
BIG-G dense	422M
BIG-G dense	1B
BIG-G dense	2B
BIG-G dense	4B
BIG-G dense	8B
BIG-G dense	27B
BIG-G dense	128B

The HELM benchmark evaluates a broader range of LLMs provided by several leading AI research organisations, including AI21 Labs, Aleph Alpha and BigScience. This range includes smaller models that are adept at specific tasks, to behemoths such as Google's PaLM with 540 billion parameters and DeepMind's Gopher with 280 billion parameters, which tackle more extensive and complex tasks. A detailed list, presented in table 2, provides a comprehensive overview of the models included in the HELM benchmark. These LLMs, which include versions of OpenAI's GPT series and various offerings from Cohere, are rigorously evaluated to highlight their analytical and generative capabilities under consistent test conditions. The HELM benchmark serves as a critical test for the current and future state of natural language understanding and generation in AI systems.

B Readability metrics

The QUANTEDA R package for managing and analysing text provides the following metrics for measuring the complexity of a text:

- **Lexical Diversity**: TTR, C, R, CTTR, U, S, K, I, D, Vm, Maas, lgV0, lgeV0, nchar.
- **Readability**: ARI, ARI.simple, ARI.NRI, Bormuth.MC, Bormuth.GP, Coleman, Coleman.C2, Coleman.Liau.ECP, Coleman.Liau.grade, Coleman.Liau.short, Dale.Chall, Dale.Chall.old, Dale.Chall.PSK, Danielson.Bryan, Danielson.Bryan.2, Dickes.Steiwer, DRP, ELF, Farr.Jenkins.Paterson, Flesch, Flesch.PSK, Flesch.Kincaid, FOG, FOG.PSK, FOG.NRI, FORCAST, FORCAST.RGL, Fucks, Linsear.Write, LIW, nWS, nWS.2, nWS.3, nWS.4, RIX, Scrabble, SMOG, SMOG.C, SMOG.simple, SMOG.de, Spache, Spache.old, Strain, Traenkle.Bailer, Traenkle.Bailer.2, Wheeler.Smith, meanSentenceLength, meanWordSyllables.

We have made a selection considering their frequency in the literature and the low collinearity between them after doing a non-exhaustive analysis of the correlation matrices for all the selected datasets. For instance, in Figure 1, we can see the correlation matrix for lexical diversity metrics in *MMLU Computer Security* task. *TTR*, which is a popular lexical diversity metric, is highly correlated with *U*, *S*, *I*, *Maas*, *lgVo* and *lgeV0*, so we can discard all these metrics that are going to provide a measure similar to TTR. This process is repeated until we are left with a representative number of metrics.

For readability metrics we use the same process. Here is an example of *MMLU Econometrics*. As the Figure is so large that reading it in the project would not be possible, it can be found here. For example, if we take a look at *Scrabble* metric, we observe that it is not highly correlated with any of the metrics, so even if it is an experimental measure created by QUANTEDA, we select it because it may provide another insight.

We finally chose TTR and K for lexical diversity and Flesch, Scrabble, FOG, SMOG.C and FORCAST for readability.

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

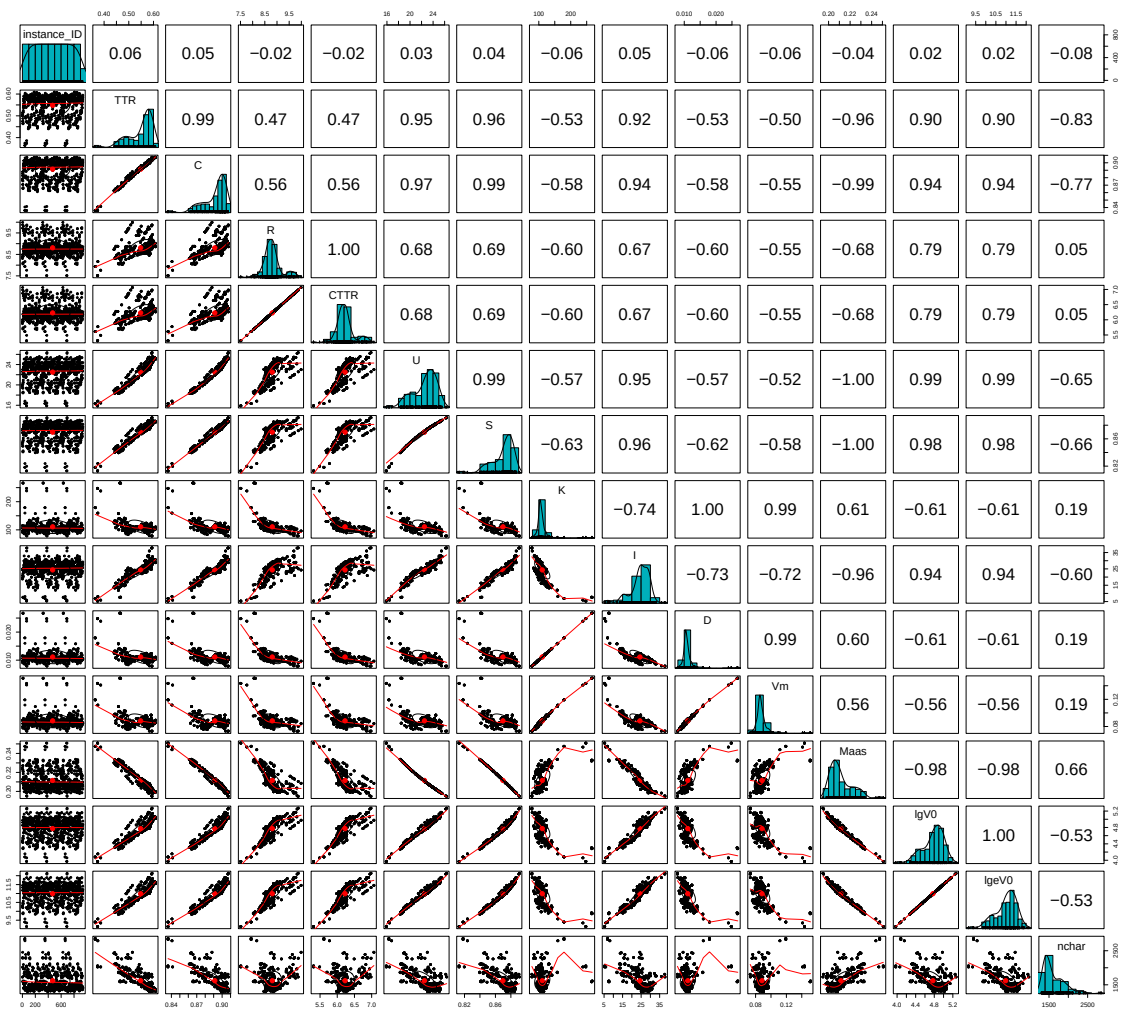

Figure 1: Correlation Matrix for Lexical Readability Metrics in *MMLU Computer Security task*

B.0.1 Lexical Diversity

In the following formulas, N refers to the total number of tokens, V to the number of types (different unique word stems) and $f_v(i, N)$ is the number of types occurring i times in a sample of length N.

- **TTR (Type-Token Ratio)**. It weights the range of vocabulary per size of the text. High TTR means less repetitive vocabulary usage. For instance, if a text contains 30 words and they are all different, its TTR would be the highest, 1.

$$TTR = \frac{V}{N}$$

- **K (Yule's K [7])**. It measures the rate at which words are repeated. Hence, it can be considered an inverse metric of lexical richness, the lower the better.

$$K = 10^4 \times \left[-\frac{1}{N} + \sum_{i=1}^{V} f_v(i, N) \left(\frac{i}{N}\right)^2 \right]$$

B.0.2 Readability

First, let us clarify that:

- n_w is the number of words.
- n_{st} equals the number of sentences.
- n_{sy} is the number of syllables
- ASL is Average Sentence Length (number of words divided by number of sentences)

Having pointed this out, the metrics are as follows

- **Flesch (Flesch's Reading Ease Score [2])**. It is a simple approach to assess the grade level of the reader. It ranges from 0 to 100. The higher the easier the text is to read.

$$Flesch = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \frac{n_{sy}}{n_w})$$

- **Scrabble (Scrabble Measure)**. Represents the mean Scrabble letter values of all words.
- **FOG (Gunning's Fog Index [3])**. The idea of this index is that short sentences written in plain English achieve a better score than long sentences written in complicated language. The ideal score for readability with the Fog index is 7 or 8. Anything above 12 is too hard for most people to read.

$$FOG = 0.4 \times (\text{ASL} + 100 \times \frac{n_{wsy \geq 3}}{n_w})$$

where $n_{wsy \geq 3}$ is the number of words with three syllables or more.

- **SMOG (SMOG (Regression Equation C) [4])**. It estimates the years of education a person needs to understand a piece of writing.

$$SMOG.C = 0.9986 \times \sqrt{Nwmin3sy \times \frac{30}{n_{\text{st}}} + 5} + 2.8795$$

Table 2: Models evaluated in HELM tasks

Creator	Model	#Parameters
AI21 Labs	J1-Jumbo v1	178B
AI21 Labs	J1-Large v1	7.5B
AI21 Labs	J1-Grande v1	17B
AI21 Labs	J1-Grande v2 beta	17B
Aleph Alpha	Luminous Base	13B
Aleph Alpha	Luminous Extended	30B
Aleph Alpha	Luminous Supreme	70B
Anthropic	Anthropic-LM v4-s3	52B
BigScience	BLOOM	176B
BigScience	BLOOMZ	176B
BigScience	T0pp	11B
BigCode	SantaCoder	1.1B
Cohere	Cohere xlarge v20220609	52.4B
Cohere	Cohere large v20220720	13.1B
Cohere	Cohere medium v20220720	6.1B
Cohere	Cohere small v20220720	410M
Cohere	Cohere xlarge v20221108	52.4B
Cohere	Cohere medium v20221108	6.1B
Cohere	Cohere command nightly	6.1B
Cohere	Cohere command nightly	52.4B
DeepMind	Gopher	280B
DeepMind	Chinchilla	70B
EleutherAI	GPT-J	6B
EleutherAI	GPT-NeoX	20B
Google	T5	11B
Google	UL2	20B
Google	Flan-T5	11B
Google	PaLM	540B
HazyResearch	H3	2.7B
Meta	OPT-IML	175B
Meta	OPT-IML	30B
Meta	OPT	175B
Meta	OPT	66B
Meta	Galactica	120B
Meta	Galactica	30B
Microsoft/NVIDIA	TNLG v2	530B
Microsoft/NVIDIA	TNLG v2	6.7B
OpenAI	davinci	175B
OpenAI	curie	6.7B
OpenAI	babbage	1.3B
OpenAI	ada	350M
OpenAI	text-davinci-003	-
OpenAI	text-davinci-002	-
OpenAI	text-davinci-001	-
OpenAI	text-curie-001	-
OpenAI	text-babbage-001	-
OpenAI	text-ada-001	-
OpenAI	code-davinci-002	-
OpenAI	code-davinci-001	-
OpenAI	code-cushman-001	12B
OpenAI	ChatGPT	-
Together	GPT-JT	6B
Together	GPT-NeoXT-Chat-Base	20B
Tsinghua	CodeGen	16B
Tsinghua	GLM	130B
Tsinghua	CodeGeeX	13B
Yandex	YaLM	100B

where $n_{wsy\geq 3}$ is the number of words with three syllables or more.

- **FORCAST (Simplified Version of FORCAST.RGL [1])**. It measures the grade level needed to read a text. It is designed to analyse technical text and it is considered ideal for multiple-choice tests, surveys and guides.

$$FORCAST = 20 - \frac{n_{wsy=1} \times 150}{(n_w \times 10)}$$

where $n_{wsy=1}$ is the number of one-syllable words.

C Linguistic Meta-features

Table 3 outlines a comprehensive selection of meta-features and their respective scales, including aspects ranging from certainty to noise levels. While the aim is to unambiguously quantify each meta-feature, some situations inherently involve subjective judgement. In particular, the level of granularity —whether the whole text, discrete paragraphs or individual sentences are assessed— and language variation can play a key role in determining the final values. Some features are defined by concrete methods, such as the proportionate occurrence of grammatical constructs, while others, particularly those derived from cognitive linguistics, may require conventional anchors (i.e., referential examples against which judgments are made).

Despite the complexities involved, this methodology allows for widespread application. Although manual annotation is sometimes required, especially where rules cannot be explicitly defined, the broad comparisons between instances match empirical standards well. It is acknowledged that certain choices regarding the definition of the different meta-features are principled decisions, aimed at demonstrating the viability of the methodology rather than establishing a definitive rubric. This element introduces scope for refinement based on further linguistic and cognitive evidence.

D Annotation postprocessing

Post-processing is required to adapt the output of the language model to the established scales, as values are sometimes outside the intended range. Out-of-range responses, including non-answers or irrelevant sentences, are standardised by assigning the minimum scale value to responses below the range and the maximum scale value to those above.

For example, we have the following output from GPT-4 for a certain example which we have to modify accordingly:

Postprocessing example

"Marcos shuld be in the lab because he has to finish his work today."
- Uncertainty: 4
- Negation: N/A → **0**
- Time: 1
- Space: 1
- Vocabulary: 0.121
- Modality: 1
- Theory of Mind: -1 → **0**
- Reasoning: 1
- Compositionality: 1
- Anaphora : 1
- Noise: 1/54 → **0.019**

Once all the values are corrected (Figure 2 and Figure 3 illustratively shows the range of values obtained for each meta-feature for the tasks *BBQ* and *Epistemic Reasoning*), we can calculate the values of the meta-features for all the instances, i.e. for the whole paragraphs.

After evaluating various aggregation methods, we found that averaging the levels of meta-features across sentences provided the most representative score for an entire text instance.

Consider a paragraph consisting of three sentences:

"Billy was so excited to win the race for class president. He had worked so hard to win, and he was so proud of himself. His opponent wasn't so happy."

In this case, there are no time-related expressions, so the 'Time' meta-feature would logically have a value of zero. Conversely, 'Negation' occurs once, and if we were to use the maximum value across sentences, the meta-feature level of the whole paragraph would be unfairly raised to one. Instead, by calculating the mean, we arrive at a Negation score of 0.33, which accurately reflects the presence and influence of negation without overstating its impact based on a single sentence. This measured approach nuances our un-

Table 3: Description of linguistic meta-features

Meta-features	Scale and Levels	Examples
Uncertainty	0: complete certainty, ... 10: complete uncertainty	"The cat is in the house": 1 "She might not do it again": 7 "He may come this afternoon": 3 "We have no clue about where it is": 8 "It is a fact that a square has four sides": 0 "It's impossible to know who will win the lottery": 10 "I'm not sure who will win the election": 8
Negation	0: no negation 1: simple negation 2: double negation 3: negation with quantification 4: very complex negation ...	"I'm a rich man" : 0 "She has never had a dog": 1 "It's untrue that all houses without windows do not have any light": 4 "I don't know what I don't know": 2 "The suspect is not in the house": 1 "The car has not been driven by anyone in the team": 3 "Never say never": 2
Time	0: no time expressions 1: simple temporal expressions 2: double temporal expressions 3: complex temporal expressions ...	"He came before noon": 1 "The house is blue" : 0 "There's a meeting every two weeks" : 3 "The train arrived ten minutes after the plane has left": 2
Space	0: no space relationships 1: simple spatial expressions 2: double spatial expressions 3: complex spatial expressions ...	"The pen was on the table": 1 "There's no room between the two cars": 2 "Tomorrow is a bank holiday": 0 "The lamp was hanging from two ropes, one attached to the ceiling and the other to the window": 5
Vocabulary	0..1: Normalised from some aggregate metric of the -log freq of words or something similar as in semantic complexity metrics.	"The ball is big": 0.1219 "Procrastination jeopardises excellence": 0.4235 "The boy must apologise": 0.198 "Ignoramus was an ultracrepidarian reposte" : 0.8324
Modality	0: no modality 1: simple modality 2: double modality ...	"The woman walked into a bar": 0 "The boy must apologise": 1 "The boy thinks we can't do it" : 3
Theory of Mind	0: no theory of mind 1: simple theory of mind 2: double theory of mind	"He came to the reception before noon": 0 "She didn't want to buy a car": 1 "The boy thinks we can't do it": 1 "The child feared his parents wanted to punish him": 2
Reasoning	0: no reasoning 1: simple reasoning 2: complex reasoning	"He tripped because of the step" : 1 "He came before noon with a bag full of presents": 0 "The grass was wet but it was sunny so someone must have watered the plant": 2
Compositionality	1...number of levels	"He came before noon": 0 "He came before she arrived": 1 "The man wearing the tall hat came before she arrived": 2 "He came before noon with a bag full of presents": 0.
Anaphora	0: no anaphora 1: simple (one possible referent) 2: complex ($>$ 1 possible referents)	"Kim thinks that he is clever": 1 "While Stuart was telling Susan the news, she laughed at him": 2
Noise	0...number of typos per character wrt to the original text with no typos	"The ball is big" : 0 "The bll isbige" : 3/13 "The boy bust apologise": 1/20

derstanding of the linguistic features of a paragraph, acknowledging their presence without overstating their pervasiveness.

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 A way of making Europe".

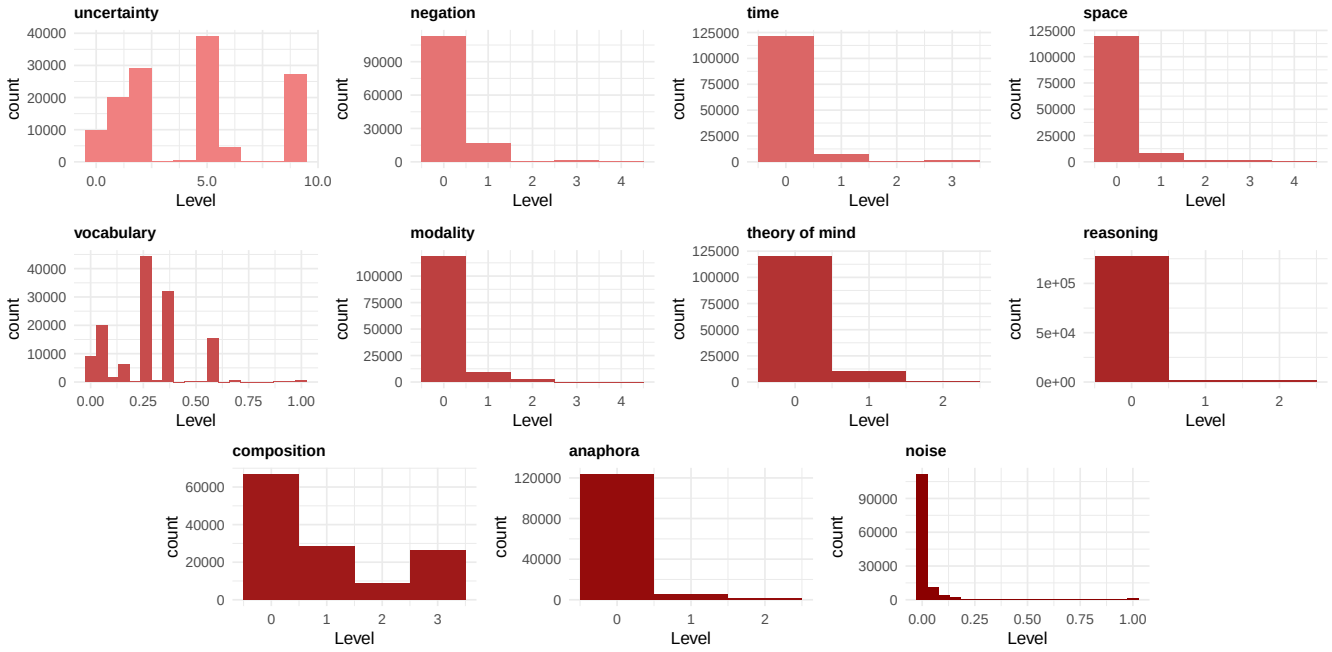

Figure 2: Resulting values for meta-features after post-processing incorrect values of *BBQ*

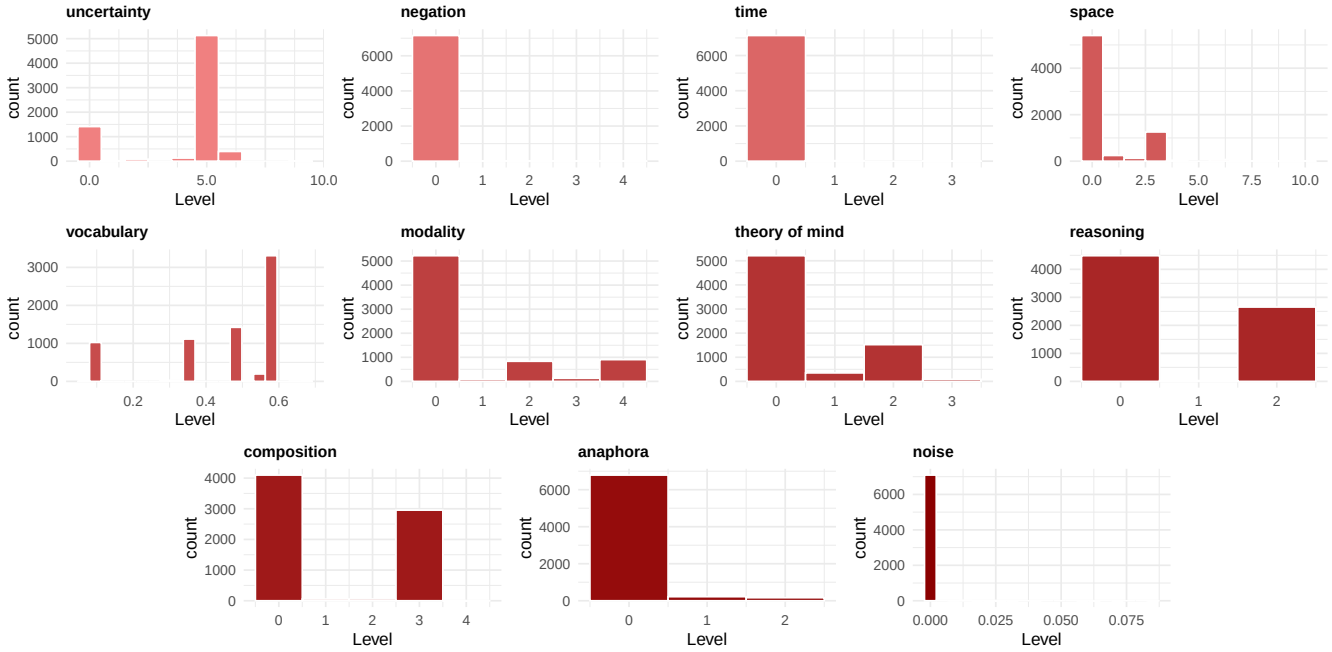

Figure 3: Resulting values for meta-features after post-processing incorrect values of *Epistemic Reasoning*

References

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

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

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

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

[5] Y. Moros-Daval, F. Martínez-Plumed, and J. Hernández-Orallo. Language task difficulty prediction through llm-annotated meta-features. In *ECAI 2024*. IOS Press, 2024.

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

[7] F. J. Tweedie and R. H. Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352, 1998.