

# Intellectual and Developmental Disabilities: The Role of LLMs in Communication

May Lutzen, Marcello Balduccini

Saint Joseph’s University, Philadelphia PA 19131, USA

**Abstract.** This paper is part of a project aimed at leveraging neuro-symbolic architectures to achieve a sophisticated interaction between humanoid robots and individuals with cognitive impairments. In our architecture, a symbolic reasoning module based on Answer Set Programming creates a sequence of appropriate activities and monitors their execution in real-time. Large Language Models (LLMs) are used to enhance the user experience in various ways. Among those, in this paper we investigate methods for using LLMs to rewording text produced by the reasoning component, while keeping length and vocabulary level consistent with the original text. The study explores the effectiveness of metrics such as length and frequency of use, in comparing the vocabulary level of the input with that of the output. We present a comparative analysis of free or moderately priced LLMs, such as GPT-3.5, Google Gemini Pro, and Claude 3 Opus. A continuous validation process is also introduced, utilizing a critic that evaluates the appropriateness of the generated output at run-time. The findings indicate that while LLMs can often produce outputs with a vocabulary level comparable to the inputs, although there are areas needing improvement, particularly in handling specific domain knowledge or less common phrases. This research contributes to the exploration of novel neuro-symbolic architectures and to the practical application of LLMs in contexts where controlled language use is essential for effective communication.

**Keywords:** Large Language Models · Neuro-Symbolic Robot Architectures · Knowledge Representation.

## 1 Introduction

Humanoid robots employed in healthcare represent a diverse and rapidly evolving research field that encompasses various areas of investigation, for example to support people with communication disabilities as described in [5]. For instance, studies have explored the use of AI-enabled devices in providing companionship and social interaction for seniors, as highlighted in [2], underscoring the potential of humanoid robots to address social isolation and promote well-being among elderly populations. Moreover, recent research [3, 6] highlighted how AI can subtly guide individuals towards specific choices with more effective *nudging* capabilities than humans, suggesting that an automated assistant might be a useful tool for guiding patients towards healthier behaviors.

This paper is part of a project stemming from a partnership with Bancroft (<https://bancroft.org>), a non-profit organization dedicated to helping individuals with autism, intellectual and developmental disabilities, brain injury, and neurological conditions. A challenge faced by service providers like Bancroft is that direct care facility staff may become overwhelmed with routine tasks, leading to low job satisfaction and retention, and hindering their performance in higher-level duties and communication with clinical staff. High turnover exacerbates these issues, as new and temporary staff lack the specific knowledge needed for effective patient care. All of these staff issues, in the long term, cause reduced quality of care thus impacting the quality of life of the individuals served. The goal of the project is to research methods at the intersection of symbolic reasoning, machine learning, and robotics that can lead to enhancing both staff satisfaction and resident care.

The project augments *Iggy*, a Pepper humanoid robot [9], with a neuro-symbolic architecture, called IR4D<sup>1</sup>, which combines the powerful knowledge representation and reasoning capabilities of Answer Set Programming (ASP) [4] with Large Language Models (LLMs).

Given information on a group of patients, the *reasoning module* builds a sequence of appropriate activities and monitors their execution, making real-time adjustments based on user responses. Whenever a menu is presented on the robot’s display, the *reasoning module* generates text that is read out by the robot to introduce the menu. This is done not only for a more natural interaction but also because some of the patients have minimal or no reading skills. Given that the robot controlled by the architecture interacts with individuals with potentially impaired cognitive skills, the text produced by the symbolic reasoning component is based on fragments that are carefully crafted by psychologists in order to ensure that the vocabulary level is suitable for the audience.

Depending on the circumstances, the same menu may be displayed multiple times in a row. In order to make the interaction more engaging for the audience and in order to make the robot sound more natural, the robot automatically produces variations of the text generated by the symbolic reasoning component. While LLM are a natural choice for generating these variations, the variations must still match the audience’s capabilities, and in particular their (receptive) vocabulary level.

In this paper, we present an initial evaluation of methods for controlling the vocabulary level of the text produced by LLMs in rewording tasks, with a specific attention to the use with individuals with cognitive impairment. We begin by considering different kinds of metrics normally associated with the vocabulary level of text, and then evaluate the use of a *continuous validation* process, in which the output produced by the LLM is vetted at runtime by a *critic* that determines if the output is acceptable. Due to considerations on practical use, in this paper we focus our evaluation on LLMs that are either free or moderately priced, but also of sufficient power and ease of use, i.e. GPT 3.5, Google Gemini

---

<sup>1</sup> IR4D stands for *Intelligent Robotics for Disabilities*

Pro, and Claude 3 Opus. In a later phase of the study, we will evaluate more powerful versions of these as well as self-hosted LLMs.

The paper is organized as follows. We begin with a discussion on methods of measuring the vocabulary of text and clarify the research questions that our analysis aims at answering. The following sections analyze metrics of progressively increasing sophistication. We conclude the paper with final considerations and a discussion on future work.

## 2 Measuring the Vocabulary Level of Text

Typically, the evaluation of the vocabulary level of text takes into account metrics such as length, frequency of use, morphological complexity and semantic complexity, as discussed, e.g., in [1, 7, 8]. Length refers to the number of characters, words, and syllables in the text, as it is conceivable that longer text might require stronger cognitive efforts. Frequency of use refers to how often words are used in everyday language. If they are less common, they might be more complex and thereby more difficult to understand. Morphological complexity refers to the complexity of a word’s internal structure and words with more complex structures might be more difficult to understand. Lastly, semantic complexity refers to the number of meanings or senses a word has. If a word has multiple meanings, reaching the desired understanding of a word can be much more complex. In this paper, we focus on the first two metrics, and use the following research questions to guide the analysis:

- How capable are the LLMs of producing output that is of comparable length to the input if instructed to do so? And how reliable is length as an indicator of vocabulary level in the context of our application domain?
- How capable are LLMs of producing output in which frequency-of-use is comparable to that of the input?
- Is there an algorithmic way of leveraging frequency-of-use as a way to improve the vocabulary level of output text?

We conduct our analysis over a benchmark set of 10 passages that covers all the categories of passages generated by the *reasoning module* during a session. Categories are represented by a uniform number of passages in the benchmark set. The passages are shown in Figure 1.

For each input passage, we produce three output passages for each LLM. We chose to produce multiple output passages in order to limit the impact of statistical anomalies on our analysis.

It is important to note that some of the passages contain the names of activities that the audience can choose from using the robot’s display. Those names are familiar to the audience. In preliminary experiments, we observed that the LLMs have a tendency to reword the activity names, often making the activities difficult to recognize, sometimes drastically altering the semantics of the names – for example, turning “a name that sound game” into “a name-based

<b>P1</b>	Hello again! It's good to see you back here at Saint Joseph's University. You may remember me from last time. My name is Pepper, and I'm a social robot here at the Innovation Center. We have more fun activities planned for you today.
<b>P2</b>	How about we take a little break now and relax? When you are ready to resume, click the button on my screen.
<b>P3</b>	Before we say goodbye, I have a little surprise for you. Here are some activities that I hope you'll like: "a name that sound game for David", "some dancing fun", and "a name that sound game". Pick from one of the boxes on my screen to start an activity!
<b>P4</b>	Guess who's back from its break? "the Frosty the Snowman sing-along". Which activity would you like to pick now? We have "the Frosty the Snowman sing-along", "David's name that sound game", and "the name that season game".
<b>P5</b>	Before we say goodbye, I have a little surprise for you. Here are some activities that I hope you'll like: "a Frosty the Snowman sing-along", "a name that sound game for David", and "a name that tune game". Pick from one of the boxes on my screen to start an activity!
<b>P6</b>	Let's give a little break to "the name that season game". Guess who's back from its break? "the emotion recognition game". Which activity would you like to pick now? We have "the emotion recognition game", "the name that tune game", and "the name that animal sound game".
<b>P7</b>	Here are a few activities for you to have fun with now. We have "Mister Sandman", "a name that tune game", and "a dance routine". Pick from one of the boxes on my screen to start an activity!
<b>P8</b>	Here are a few activities for you to have fun with now. We have "a Frosty the Snowman sing-along", "an opportunity to chat with Pepper, and "a name that sound game for David". Pick from one of the boxes on my screen to start an activity!
<b>P9</b>	Here are a few activities for you to have fun with now. We have "a name that sound game", "a Rudolph the Red Nosed Reindeer sing-along", and "a name that animal sound game". Pick from one of the boxes on my screen to start an activity!
<b>P10</b>	Let's give a little break to "the Frosty the Snowman sing-along". Which activity would you like to pick now? We have "chat with Pepper", "dancing fun", and "David's name that sound game".

**Fig. 1.** Benchmark set used in the analysis

sound game”, and even rewording “an emotion recognition game” into “a mind-reading game”. To our surprise, we were only partially successful at designing prompts that directly prevented the LLMs from rewording the activity names, and resorted to an algorithmic approach for preserving them. In our algorithmic preservation approach, the activity names are replaced by strings that the LLMs do not appear to reword, such as “ $X1X$ ”. After receiving the LLM’s output, they are then converted back to produce the final output. While in our analysis we considered both kinds of approaches to activity name preservation, in this paper we focus on the analysis of the passages obtained through the algorithmic preservation approach. Additionally, due to space considerations, in the sections that follow we will be providing only excerpts of some of the tables. The complete tables can be found at <https://g34r.short.gy/dataset>.

### 3 Length-Based Metrics

Length of text is a readily-usable metric in the context of LLMs, as phrases such as “of length comparable to the input” can be easily included in the prompt to an LLM. In this section, we evaluate the output produced by the LLMs under consideration when the prompt instructs the LLM to constrain the length of the output. Specifically, we use the prompt:<sup>2</sup>

*“Take what I say after the new-line character and paraphrase it. The new text should have a similar length to the one in input as well as the same tone and language.”*

The complete list of the output passages produced by the LLMs can be found at the link provided earlier. We provide sample output passages – specifically for P10 – in Figure 2.

In this section, we compare the various LLMs in terms of number of characters, words, and syllables of the output w.r.t. to those of the input.

For every metric and LLM, we calculate average change (i.e., difference between input length and output length) for every passage, as well average change across all passages. The results of the analysis on length in characters are shown in Figure 3. The analysis of the average change, per LLM, on each input passage shows a clear trend. GPT has the lowest average change, followed by Gemini and then Claude, which has the highest average change. This is also the case for the aggregate average change on all passages, which is shown in the 3 bottom rectangles for each metric, where GPT has 0% average change, Gemini 10%, and Claude 44%. The results from the analysis of length measured in characters is confirmed by looking at the average change in length measured by word count

<sup>2</sup> One might argue that more sophisticated ways of phrasing the prompt should be considered. While that is indeed possible, we intentionally stay away from the route of prompt engineering and focus on analyzing what LLMs are capable of producing given a reasonably precise prompt that an average human would likely be able to understand and act upon.

LLM	Output
GPT	Let's take a breather and switch things up for "the Frosty the Snowman sing-along". Which option do you fancy next? We've got "chat with Pepper", "dancing fun", and "David's name that sound game" on the table.
GPT	Let's take a brief pause for "the Frosty the Snowman sing-along". What would you like to do next? Your options are "chat with Pepper", "dancing fun", and "David's name that sound game".
GPT	Let's take a short pause for "the Frosty the Snowman sing-along". What activity do you feel like doing next? We've got options like "chat with Pepper", "dancing fun", and "David's name that sound game".
GEMINI	After a little break, "the Frosty the Snowman sing-along" needs a moment to unwind. What new pursuit do you suggest for "the Frosty the Snowman sing-along"? "chat with Pepper", "dancing fun", and "David's name that sound game" are all possibilities.
GEMINI	Let's give "the Frosty the Snowman sing-along" a quick breather. How about trying out "chat with Pepper"? Or perhaps "dancing fun" or "David's name that sound game" might be a better fit for you right now?
GEMINI	Let's put "the Frosty the Snowman sing-along" on hold for a bit. What would you like to do next? We can choose from "chat with Pepper", "dancing fun", or "David's name that sound game".
CLAUDE	Here is the paraphrased text with the specified variables unchanged: How about we allow "the Frosty the Snowman sing-along" to rest for a bit? What would you prefer to do instead? The options available are "chat with Pepper", "dancing fun", and "David's name that sound game".
CLAUDE	Here is the paraphrased text with variables unchanged: How about we allow "the Frosty the Snowman sing-along" to rest for a bit? What would you prefer to do instead? The options available are "chat with Pepper", "dancing fun", and "David's name that sound game".
CLAUDE	Here is the paraphrased text with the specified variables unchanged: How about we allow "the Frosty the Snowman sing-along" to rest for a bit? What would you prefer to do instead? The options available are "chat with Pepper", "dancing fun", and "David's name that sound game".

Fig. 2. Output passages produced for passage P10

Passage #	LLM	Length of input	Length of output	Change in length	Average change for each passage	Average change on all passages	Word count of input	Word count of output	Change in word count	Average change for each passage	Average change on all passages	Syllable count of input	Syllable count of output	Change in syllables	Average change for each passage	Average change on all passages
P9	GPT	244	246	1%			47	46	-2%			69	71	3%		
P9	GEMINI	244	232	-5%			47	43	-9%			69	68	-1%		
P9	CLAUDE	244	531	118%			47	92	96%			69	149	116%		
P9	GPT	244	210	-14%			47	37	-21%			69	57	-17%		
P9	GEMINI	244	244	0%			47	42	-11%			69	68	-1%		
P9	CLAUDE	244	296	21%			47	48	2%			69	86	25%		
P9	GPT	244	257	5%	-3%		47	47	0%	-8%		69	72	4%	-3%	
P9	GEMINI	244	273	12%	2%		47	50	6%	30%		69	77	12%	3%	
P9	CLAUDE	244	386	58%	66%		47	62	32%	43%		69	107	55%	65%	
P10	GPT	198	209	6%			33	36	9%			48	51	6%		
P10	GEMINI	198	249	26%			33	40	21%			48	64	33%		
P10	CLAUDE	198	277	40%			33	46	39%			48	74	54%		
P10	GPT	198	185	-7%			33	32	-3%			48	46	-4%		
P10	GEMINI	198	205	4%			33	36	9%			48	50	4%		
P10	CLAUDE	198	263	33%			33	44	33%			48	70	46%		
P10	GPT	198	202	2%	0%	0%	33	34	3%	3%	-4%	48	52	8%	3%	2%
P10	GEMINI	198	185	-7%	8%	10%	33	34	3%	11%	5%	48	45	-6%	10%	10%
P10	CLAUDE	198	277	40%	38%	44%	33	46	39%	37%	34%	48	74	54%	51%	49%

**Fig. 3.** Change in length measured in characters, word count and syllables

where the average change on all the different passages by LLM for GPT is -4%, Gemini shows 5% average change and Claude shows 34% average change.

While length and average change in length might show how well a large language model is able to follow directions from a prompt, our analysis highlighted that, at least for our target population, the length of the text is not a good indicator of the vocabulary level. Consider for instance the input text:

*Before we say goodbye, I have a little surprise for you. Here are some activities that I hope you'll like: "a name that sound game for David", "some dancing fun", and "a name that sound game". Pick from one of the boxes on my screen to start an activity! "*

The output created by Claude is:

*Hold your horses! Before we part ways, I've got a treat in store. I've handpicked some cracking activities that I reckon you'll dig: "a name that sound game for David", "some dancing fun", and "a name that sound game". Go ahead and click on any of the boxes on your screen to get the party started!*

The vocabulary level of phrases such as "hold your horses", "handpicked some cracking activities" and "I reckon you'll dig" is *significantly* different from "I have a little surprise" and "I hope you'll like", even though the change in number of characters (13%) and in number of words (12%) might be considered modest by comparison.

When it comes to syllables, Figure 3 shows that the overall average change in number of syllables resembles follows trends similar to those of the number of characters and words. GPT has the overall smallest average change with only 2% from input to output, Gemini is second with only 10% overall average change from input to output and Claude performed the worst with 49% overall average change from input to output. The conclusion of this analysis thereby shows that GPT is best at following directions in the prompt given to it specifically about desired length in a paraphrased output whereas Gemini is slightly worse but can still be considered acceptable as it within a 10% change. However, in many of the 30 Claude cases in our analysis, the LLM includes part of the prompt in its output, such as 'with the specified variables'. It should therefore be noted that one reason for Claude's great positive change in length is due to the fact that Claude tends to include part of the prompt in the output thereby increasing the length, however this also shows how it is unable to follow directions from the prompt.

However, the number of syllables does not appear to have a strong relationship with the vocabulary level, at least for the types of passages and population considered here. Consider one of the outputs produced by Claude for P5:

- **Input:** *Before we say goodbye, I have a little surprise for you. Here are some activities that I hope you'll like: "a Frosty the Snowman sing-along", "a name that sound game for David", and "a name that tune game". Pick from one of the boxes on my screen to start an activity!*



- **Output:** *Before we part ways, I’ve got a special treat in store for you. Take a look at these fun options I’ve prepared: “a Frosty the Snowman sing-along”, “a name that sound game for David”, and “a name that tune game”. Simply select one of the boxes displayed to embark on an exciting activity!*

Here, phrases such as “part ways”, “special treat in store for you” and “embark on” have a significantly different vocabulary level than “goodbye”, “little surprise” and “to start”.

In conclusion, while the change in length between input and output can tell us how well an LLM is at following directions related to the length of the output – and is useful in preventing the output from being boring or wordy – it does not help ensure that the vocabulary level of the output matches that of the input.

## 4 Frequency-of-Use Metric

Of the other possible options for estimating the vocabulary level of a passage, the frequency of use (abbreviated **uf** in this paper) of its words appears to be a practically viable metric. The intuition behind this metric is simple: frequency of use refers to how often certain words are used in a corpus of text that is used as a reference; if words are less common, it is conceivable that they are more complex or difficult to understand.

The minimum frequency of use (abbreviated **min uf**) across all the words of a passage seems particularly useful in our application domain. After all, we are looking for text that, when spoken by the robot, can be *promptly* understood by our target population. Even a single word that is outside of the receptive vocabulary of the audience can be sufficient to prevent them from promptly understanding the passage. It must be noted, however, that it is not sufficient to consider the **uf** of individual words, as phrases, especially if idiomatic, may have a meaning that is different from that of their constituent words and a frequency of use that is also different from of the constituent words. Consider for example “let the fun begin”. While the individual words of this phrase are fairly common, the specific combination is not as frequently used as, for example “let’s have fun.” For this reason, we adopt the following approach.

Let  $\tau(p)$  denote the parse tree associated with passage  $p$  (if multiple parse trees exist, we assume that a parse tree has been arbitrarily selected). For a node  $n \in \tau(p)$ , let  $uf(n)$  denote the frequency of use of the text associated with  $n$ .  $uf(n)$  is undefined if no frequency of use is available for the text.

**Definition 1.** A node  $n$  of  $\tau(p)$  is a **uf-top** node if:

- $uf(n)$  is defined, and
- $uf(n')$  is undefined for every ancestor  $n'$  of  $n$  in  $\tau(p)$ .

We can now give the following:

**Definition 2.** Given a passage  $p$ :

- The `min uf` of  $p$ , written  $\text{min-uf}(p)$ , is

$$\text{min-uf}(p) = \sum_{n \in \tau(p)} \text{uf}(n)$$

- The `min-uf` phrase of  $p$  is the fragment of text associated with a `uf-top` node of  $\tau(p)$  that has minimum `uf`. (For simplicity, if multiple such nodes exist, we select the first node in leftmost traversal of the parse tree.)

For the analysis that follows, the parse trees were generated using the *spacy* library<sup>3</sup> with the *en\_core\_web\_lg* model. The `uf` of phrases has been obtained from the *absTotalMatchCount* metric of the NGRAMS tool<sup>4</sup> with default settings<sup>5</sup>. NGRAMS uses the *Google Books Ngram Dataset v3* as the source of ngram data.

At least at first sight, the `min uf` appears to provide meaningful indication of the vocabulary level of passages. For example, the phrase “on my screen” which one might consider fairly simple to understand, has a comparatively high `min uf` of 33090, while phrases such as “our prior encounter” and “infringing on any copyrights,” which arguably belong to a higher vocabulary level, have a comparatively low `min uf` of 72 and 42 respectively.

So, the question is: how capable are LLMs of producing output whose `min uf` is comparable to the `min uf` of the input given our prompt?

To answer the question, we compared the `min uf` of our input and output passages. However, one should notice that activity names deserve some attention. In fact, it so happens that some activity names are rather infrequent in the corpus used for the calculation of `min uf`, even though they are promptly understood by the audience, causing them to have a low use frequency that may bias the evaluation. Since the rewording process ensures that the activity names are kept verbatim in the output, in many cases those are also the *min-uf phrases* of the output.

For this reason, we disregard the activity names in the calculation of the `min uf`, leading to the comparison outlined in Figure 4. We observe that in 50% of cases the `min uf` of the output was worse than that of the input. That is, in only 50% of the cases the passage produced by the LLM was satisfactory in terms of `min uf`.

It is worth noting that a potential issue with the ngrams used for estimating frequency of use. As seen from Figure 4, in a number of cases in which the `min uf` of the output is greater than or equal to the `min uf` of the input, the *min-uf phrase* of the input is “Which activity would you like.” For some reason, that phrase is assigned a rather low `uf`, which causes most outputs to have a better `min uf` than the input passage – even in cases where the *min-uf phrase* of the output is “with the specified variables”, which intuitively seems to be at a higher

<sup>3</sup> <https://spacy.io>

<sup>4</sup> <https://ngrams.dev/>

<sup>5</sup> It is certainly possible that different settings may affect the results of our analysis, particularly if one selects specific year ranges for the frequency of the ngrams. We defer an evaluation of the impact of these settings to a later phase of our study.

LLM	Input's min-uf phrase	Output's min-uf phrase	Revised min-uf input	Revised min-uf output	Change in min uf input to min uf of output	Greater or Less
GPT	to start an activity'	on my display'	2471	1746	-29%	Less
GEMINI	to start an activity'	'on my display'	2471			
CLAUDE	to start an activity'	'activities to enjoy'	2471	3143	27%	Greater
GPT	to start an activity'	available on the screen'	2471	3623	47%	Greater
GEMINI	to start an activity'	'to dive into the fun'	2471			
CLAUDE	to start an activity'	'on your chosen activity'	2471			
GPT	to start an activity'	on my display'	2471	1746	-29%	Less
GEMINI	to start an activity'	'to start an activity'	2471	2471	0%	Greater
CLAUDE	to start an activity'	an activity of your choice'	2471	185	-93%	Less
GPT	to start an activity'	""s explore"	2471	17961	627%	Greater
GEMINI	to start an activity'	To start an activity,'	2471	398	-84%	Less
CLAUDE	to start an activity'	on my interface'	2471	111	-96%	Less
GPT	to start an activity'	""s have some fun"	2471	2384	-4%	Less
GEMINI	to start an activity'	are all available.'	2471	15684	535%	Greater
CLAUDE	to start an activity'	'Ready for some excitement?'	2471			
GPT	to start an activity'	of entertaining options'	2471	131	-95%	Less
GEMINI	to start an activity'	sure to keep you entertained'	2471	94	-96%	Less
CLAUDE	to start an activity'	'some entertaining activities'	2471			
GPT	to start an activity'	to begin an activity'	2471	1130	-54%	Less
GEMINI	to start an activity'	of these exciting activities'	2471	67	-97%	Less
CLAUDE	to start an activity'	some exciting activities'	2471	206	-92%	Less
GPT	to start an activity'	the Red Nosed Reindeer'	2471	3148	27%	Greater
GEMINI	to start an activity'	the Red Nosed Reindeer'	2471	3148	27%	Greater
CLAUDE	to start an activity'	your desired activity'	2471	68	-97%	Less
GPT	to start an activity'	some fun options'	2471	85	-97%	Less
GEMINI	to start an activity'	to dive into the fun'	2471	49	-98%	Less
CLAUDE	to start an activity'	displayed on my screen'	2471	210	-92%	Less
GPT	Which activity would you like'	fancy next?'	57	52	-9%	Less
GEMINI	Which activity would you like'	What new pursuit'	57	69	21%	Greater
CLAUDE	Which activity would you like'	with the specified variables'	57	114	100%	Greater
GPT	Which activity would you like'	Your options are'	57	37572	65816%	Greater
GEMINI	Which activity would you like'	a better fit for you'	57	1278	2142%	Greater
CLAUDE	Which activity would you like'	the paraphrased text'	57	477	737%	Greater
GPT	Which activity would you like'	like doing next'	57	142	149%	Greater
GEMINI	Which activity would you like'	""s put"	57	73641	129095%	Greater
CLAUDE	Which activity would you like'	with the specified variables'	57	114	100%	Greater

Fig. 4. Change in min uf from revised input to revised output (excerpt)

vocabulary level. As we discuss later in this paper, we hypothesize that better quality ngram data may reduce this type of noise.

## 5 A Vocabulary Level Critic: an Evaluation

In the previous section, we saw how LLMs have some capability of producing output of comparable vocabulary level to the input *as estimated via  $\mathbf{uf}$* , although the performance of LLMs is less than ideal. In this section, we consider the possibility of algorithmically filtering the output of the LLMs based on this metric, and evaluate whether  $\mathbf{uf}$  provides indeed a reliable filtering mechanism.

For our evaluation, we consider a simple filtering mechanism based on a continuous validation of the output of the LLM being used. That is, we can leverage  $\min \mathbf{uf}$  to develop a *critic* that compares the  $\min \mathbf{uf}$  of the output with that of the input. If the  $\min \mathbf{uf}$  of the output is lower than that of the input, the critic rejects the output, and the architecture may decide to either have the LLM produce a new output, or use the input directly. In practice, the decision will be likely based on timeliness considerations: given the statistical nature of LLMs, there is no telling how many times the output will need to be regenerated before something of acceptable quality is produced, and long delays before the robot speaks might make for an unnatural experience.

Our goal in this section is to answer the question: *suppose the critic rejects the output produced by the LLM; based on human judgement, is that a good decision or a bad decision?* This evaluation will give us some insights both on the reliability of  $\min \mathbf{uf}$  as a filtering metric, and if the critic as a practically viable algorithmic filtering method.

The results of this evaluation are summarized in Figure 5, where a judgement of “good decision” is indicated as a true positive, and a bad decision as a false positive. At this early stage of the project, we relied on our own judgment to determine whether a decision was good or bad. In the future, we plan to conduct a more rigorous evaluation with independent test subjects. In spite of the simplicity of our validation method, we believe that the evaluation provided some interesting results. As the figure illustrates, out of 53 cases flagged by the critic, 33 are true positives, which means that in 62% of cases,  $\min \mathbf{uf}$  is a reliable indicator of whether the vocabulary level of the output produced by an LLM is acceptable. Figure 6 provides a breakdown of the information for each LLM. For GPT, the true positive rate is 61%, for Gemini it is 79%, and for Claude it is 52%. One consideration to make is that the generation of the parse tree, and possibly also the ngram data, appear to introduce some noise in the filtering process. For example, some of the cases marked as false positives during our validation are due to the string “’s have some fun” being the *min-uf phrase* in the output, and being associated with a low  $\mathbf{uf}$ . Those cases were marked as false positives during our validation because it is clear that the string refers to “let’s have some fun”, and the vocabulary level of that phrase is low. We speculate that, had the parse tree correctly extracted the string “let’s have some fun”, its  $\mathbf{uf}$  might have been higher and thus the output would not have been rejected on

LLM	Input's min-uf phrase	Output's min-uf phrase	Output passage has lower min uf	Human Validation	Produced by GPT	Produced by Gemini	Produced by Claude
GPT	at the Innovation Center'	'lined up for you today'	1	true positive	true positive		
GEMINI	at the Innovation Center'	to Saint Joseph's University	1	false positive		false positive	
CLAUDE	at the Innovation Center'	to Saint Joseph's University	1	false positive			false positive
GPT	at the Innovation Center'	to Saint Joseph's University	1	false positive	false positive		
GEMINI	at the Innovation Center'	to Saint Joseph's University	1	false positive		false positive	
CLAUDE	at the Innovation Center'	to Saint Joseph's University	1	false positive			false positive
GPT	at the Innovation Center'	to Saint Joseph's University	1	false positive	false positive		
GEMINI	at the Innovation Center'	'the vibrant campus'	1	true positive		true positive	
CLAUDE	at the Innovation Center'	to Saint Joseph's University	1	false positive			false positive
GPT	'How about we take'	'on my display'	1	true positive	true positive		
GEMINI	'How about we take'	'refreshed and ready to continue'	1	true positive		true positive	
CLAUDE	'How about we take'	'variables preserved'	1	true positive			true positive
GEMINI	'How about we take'	'energized and ready to continue'	1	true positive		true positive	
CLAUDE	'How about we take'	'the paraphrased text'	1	true positive			true positive
GPT	'How about we take'	'on my interface'	1	true positive	true positive		
CLAUDE	'How about we take'	'on my display'	1	true positive			true positive
GPT	Before we say goodbye'	Before we part ways'	1	true positive	true positive		
GEMINI	Before we say goodbye'	some exciting activities'	1	false positive		false positive	
CLAUDE	Before we say goodbye'	displayed on my screen'	1	false positive			false positive
GPT	Before we say goodbye'	Before we bid farewell'	1	true positive	true positive		
GEMINI	Before we say goodbye'	Before we part ways'	1	true positive		true positive	
CLAUDE	Before we say goodbye'	displayed on my screen'	1	false positive			false positive
GPT	Before we say goodbye'	a few fun activities'	1	false positive	false positive		
GEMINI	Before we say goodbye'	"that you'll dig"	1	true positive		true positive	
CLAUDE	Before we say goodbye'	on my interface'	1	true positive			true positive
GPT	Before we say goodbye'	to begin an activity'	1	false positive	false positive		
GEMINI	Before we say goodbye'	Before we part ways'	1	true positive		true positive	
CLAUDE	Before we say goodbye'	a few fun things'	1	false positive			false positive
GPT	Before we say goodbye'	Before we bid adieu'	1	true positive	true positive		
CLAUDE	Before we say goodbye'	of these fun activities'	1	false positive			false positive
GPT	Before we say goodbye'	Before we bid farewell'	1	true positive	true positive		
GEMINI	Before we say goodbye'	to start the fun'	1	true positive		true positive	
CLAUDE	Before we say goodbye'	Before we part ways'	1	true positive			true positive
CLAUDE	'Which activity would you like'	'your next pursuit'	1	true positive			true positive
GPT	to start an activity'	on my display'	1	true positive	true positive		
GEMINI	to start an activity'	'to dive into the fun'	1	true positive		true positive	
CLAUDE	to start an activity'	'on your chosen activity'	1	false positive			false positive
GPT	to start an activity'	on my display'	1	true positive	true positive		
CLAUDE	to start an activity'	'some entertaining activities'	1	true positive			true positive
CLAUDE	to start an activity'	on my interface'	1	true positive			true positive
GPT	to start an activity'	""s have some fun"	1	false positive	false positive		
CLAUDE	to start an activity'	'Ready for some excitement?'	1	false positive			false positive
GPT	to start an activity'	of entertaining options'	1	true positive	true positive		
GEMINI	to start an activity'	sure to keep you entertained'	1	true positive		true positive	
CLAUDE	to start an activity'	'some entertaining activities'	1	true positive			true positive
GPT	to start an activity'	to begin an activity'	1	false positive	false positive		
GEMINI	to start an activity'	of these exciting activities'	1	true positive		true positive	
CLAUDE	to start an activity'	some exciting activities'	1	true positive			true positive
CLAUDE	to start an activity'	your desired activity'	1	true positive			true positive
GPT	to start an activity'	some fun options'	1	false positive	false positive		
GEMINI	to start an activity'	to dive into the fun'	1	true positive		true positive	
CLAUDE	to start an activity'	displayed on my screen'	1	false positive			false positive
GPT	Which activity would you like'	fancy next?'	1	true positive	true positive		

Fig. 5. Human validation of revised min uf phrases

Human validation results	Overall	GPT	Gemini	Claude
True positive	33	11	11	11
False positive	20	7	3	10
Total	53	18	14	21
<b>True positive %</b>	<b>62%</b>	<b>61%</b>	<b>79%</b>	<b>52%</b>
<b>False positive %</b>	<b>38%</b>	<b>39%</b>	<b>21%</b>	<b>48%</b>

Fig. 6. Comparative results of the human validation process

its account. (A similar result would have been obtained if the ngram data had not included the entry “’s have some fun”.) That is, we hypothesize that better quality of the parsing process and of ngram data might lead to a higher accuracy of the filtering process.

The data also highlights another interesting aspect. In a certain number of cases marked as false positives during the human validation, the *min-uf phrases* in question contain place names related to the location of the robot, such as “Innovation Center” and “Saint Joseph’s University.” Clearly, those names occur infrequently in the text corpus used for determining the *uf*, but are well known to the audience. The results obtained after excluding those place names from the calculation of the *min uf* are illustrated in Figure 7. With this refinement,

Human validation results	Overall w/o Place	GPT	Gemini	Claude
True positive	31	10	10	11
False positive	13	5	1	7
Total	44	15	11	18
<b>True positive %</b>	<b>70%</b>	<b>67%</b>	<b>91%</b>	<b>61%</b>
<b>False positive %</b>	<b>30%</b>	<b>33%</b>	<b>9%</b>	<b>39%</b>

Fig. 7. Comparative results of the human validation process excluding place names

the percentage of true positives is substantially higher, standing at 70%, with a 67% for GPT, 91% for Gemini, and 61% for Claude.

## 6 Conclusions and Future Work

In this paper, we presented an initial evaluation of methods for controlling the vocabulary level of the text produced by LLMs in rewording tasks, with specific attention to the use with individuals whose cognitive capabilities are impaired. We considered length-based metrics as well as frequency of use, and defined an approach for identifying which nodes of the parse tree should be considered when measuring frequency of use. While there is space for improvement, we hope our results provide an informative initial account on the topic. For practical reasons,

at this stage we considered a set of free or moderately-priced LLMs. In the future, we plan to evaluate more powerful versions of those LLMs as well as self-hosted LLMs. We will also conduct a human validation with independent test subjects, investigate morphological and semantic complexity measures, and explore the ability of more sophisticated prompt designs to ensure a consistent vocabulary level.

*Acknowledgments.* This paper was made possible in part by funding provided by Bancroft. The passages evaluated in this paper were assembled from fragments prepared by Sara Girotto.

## References

1. Readability and the Common Core’s Staircase of Text Complexity (Aug 2012), <https://textproject.org/paper/readability-and-the-common-core-staircase-of-text-complexity/>
2. 6ABC – Action News Philadelphia: Seniors using AI robots to combat loneliness (<https://rebrand.ly/seniors-AI>) (2023), [6abc.com/seniors-ai-robot-companionship/14154002/](https://6abc.com/seniors-ai-robot-companionship/14154002/)
3. Ganapini, M.B., Fabiano, F., Horesh, L., Loreggia, A., Mattei, N., Murugesan, K., Pallagani, V., Rossi, F., Srivastava, B., Venable, B.: Value-based fast and slow ai nudging (2023)
4. Gelfond, M., Lifschitz, V.: Classical Negation in Logic Programs and Disjunctive Databases. *New Generation Computing* **9**, 365–385 (1991)
5. Khan, F., Wu, Y., Dray, J., Hemsley, B., Kocaballi, A.B.: Conversational agents to support people with communication disability: A co-design study with speech pathologists. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems CHI EA ’25*, 1–9 (Apr 2025). <https://doi.org/10.1145/3706599.3719800>, <https://dl.acm.org/doi/10.1145/3706599.3719800>
6. Mele, C., Russo Spena, T., Kaartemo, V., Marzullo, M.L.: Smart nudging: How cognitive technologies enable choice architectures for value co-creation. *Journal of Business Research* **129**, 949–960 (2021). <https://doi.org/https://doi.org/10.1016/j.jbusres.2020.09.004>, <https://www.sciencedirect.com/science/article/pii/S0148296320305786>
7. Rayner, K., Duffy, S.A.: Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* **14**(3), 191–201 (May 1986). <https://doi.org/10.3758/BF03197692>, <https://doi.org/10.3758/BF03197692>
8. Shardlow, M., Evans, R., Zampieri, M.: Predicting lexical complexity in English texts: the Complex 2.0 dataset. *Language Resources and Evaluation* **56**(4), 1153–1194 (Dec 2022). <https://doi.org/10.1007/s10579-022-09588-2>, <https://doi.org/10.1007/s10579-022-09588-2>
9. Softbank Robotics: Meet Pepper: The robot built for people (<https://us.softbankrobotics.com/pepper>) (2023), <https://us.softbankrobotics.com/pepper>