Can LLMs analyze language complexity?

Birger Moëll
KTH
bmoell@kth.se

Fredrik Sand Aronsson
Karolinska Institutet
fredisk.sand@ki.se

Johan Boye
KTH
jboye@kth.se

Abstract

Large Language Models (LLMs) have demonstrated impressive language generation capabilities but still exhibit challenges in analytical tasks requiring precise calculations, such as readability assessment and structural parsing. This paper examines the performance of three state-of-the-art LLMs—Gemini Advanced from Google, GPT-4o, and ChatGPT-o1-preview from OpenAI—on two specific tasks related to language complexity: the computation of the LIX readability metric and the Average Dependency Distance (ADD). Using a set of Swedish high school and university-level essays, we evaluate the models' ability to compute LIX and perform dependency parsing, comparing their results to established gold standards. Our findings indicate that while all models show potential, ChatGPT-o1-preview performs most consistently, closely aligning with the gold standard in both tasks. These results suggest that, with further refinement, LLMs can effectively handle tasks combining linguistic complexity and mathematical reasoning.

1 Introduction

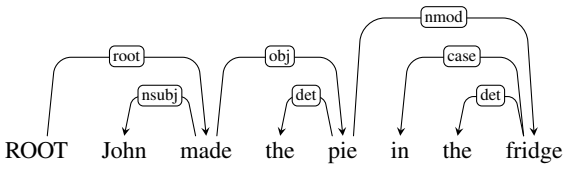
Large Language Models (LLMs) are being developed and improved at breathtaking speed. Until very recently, it seemed that models like ChatGPT could generate language in an impressive fashion, but often failed at mathematical and analytical tasks where a single correct answer was required (Wang et al., 2024a; Li et al., 2024; Mittal et al., 2024; Wang et al., 2024b).

This paper investigates how well three state-of-the-art models perform on two different analytical tasks related to language complexity: (1) computation of the LIX readability metric, and (2) dependency parsing and the computation of the Average Dependency Distance metric. These tasks are interesting as they assess the mathematical capabilities of a model (in the case of LIX), and its structural reasoning capabilities (in the case of dependency parsing). LIX, in particular, is interesting since it requires counting the number of letters in tokens, a task that should be difficult as tokens are translated to numerical identifiers in the LLM, and information about the internal structure of words should be lost.

We investigate the three models Gemini Advanced (from Google), GPT4o, and ChatGPT-o1-preview (from OpenAI), the latter model released on Sept. 12, 2024. Each of the models is given the same prompt for each of the tasks, and the output is compared to the gold standard. All prompts are found in the Appendix.

2 Language complexity metrics

Average Dependency Distance (ADD), suggested by Liu (2008), calculates the distance between each word (except the root word) and its head in a dependency tree (Kübler et al., 2009). For example, in the dependency tree for the sentence "John made the pie in the fridge" below, the distance from "fridge" to "pie" is 3.



The ADD is typically between 1.8 and 3.6 over a range of different languages (Liu, 2008).

LIX, a readability index suggested by Björnsson (1968), is computed as $A/B + 100C/A$, where A is the number of words in the text, B the number of sentences, and C is the number of words longer than six letters. Higher LIX values indicate a more advanced text: Typically LIX<30 is considered an easy text, whereas LIX>50 is advanced, and LIX>60 very advanced (e.g., research papers).

3 Method

We randomly selected 5 university-level essays, called "u_1" to "u_5" below, and 5 high-school-level essays, called "h_1" to "h_5" below[1]. All essays were written in Swedish before 2018 (to ensure that the author had not used any generative AI writing tool). From each essay, we randomly selected one paragraph (80-120 words) for the LIX calculations, and one sentence (17-42 tokens with an average of 27.8 tokens) for the dependency parsing experiment. All paragraphs and sentences can be found in the Appendix.

We computed the ground-truth LIX values using the LIX calculator at `https://www.lix.se`. The gold standard dependency trees were produced using the Stanza library[2]. We then asked the three models to compute the LIX score of each paragraph, and then to analyze our selected sentences and print their dependency trees. We instructed the models to print one word per row, on the following format:

```
1, Han, 2, 1
```

i.e., the word index, the word itself, the index of the headword, and the dependency distance between the word and its head. The exact prompts can be found in the appendix.

3.1 Models

We evaluated the following models, Gemini, GPT4-o, ChatGPT-o1-preview in order to compare performance of several state of the art models.

4 Results

4.1 Lix

Gemini (w = number of words, s = number of sentences, lw = number of long words)

Text	w	s	lw	LIX
u_1	64	7	11	26
u_2	57	6	12	31
u_3	78	10	13	24
u_4	87	7	20	35
u_5	92	8	19	32
h_1	60	5	13	34
h_2	54	5	8	26
h_3	64	5	8	25
h_4	70	5	12	31
h_5	76	8	16	30

GPT-4-o

Text	w	s	lw	LIX
u_1	42	7	3	13
u_2	58	5	8	25
u_3	73	6	9	24
u_4	91	6	12	28
u_5	119	6	12	30
h_1	68	5	6	22
h_2	53	5	3	16
h_3	58	6	3	15
h_4	91	7	9	23
h_5	97	7	7	21

ChatGPT–o1 preview

Text	w	s	lw	LIX
u_1	78	6	12	28
u_2	65	5	13	33
u_3	87	7	17	32
u_4	117	5	19	40
u_5	123	5	28	47
h_1	80	5	16	36
h_2	54	5	16	40.43
h_3	74	5	20	41.83
h_4	91	5	22	42.38
h_5	92	6	25	42.5

gold standard

Text	w	s	lw	LIX
u_1	76	7	12	27
u_2	65	5	12	31
u_3	86	7	18	33
u_4	117	5	19	40
u_5	123	5	33	51
h_1	80	5	16	36
h_2	55	5	16	38
h_3	74	5	21	43
h_4	91	5	23	43
h_5	90	7	21	36

Average differences to gold standard

Model	w	s	lw	LIX
Gemini	13.5	1.0	6.1	8.4
GPT 4-o	12.1	0.6	11.9	16.1
ChatGPT–o1	0.6	0.2	1.2	1.9

ChatGPT-o1 clearly outperforms the other models with only minor differences from the gold standard. This shows that the model can likely be used to calculate Lix scores directly.

4.2 Average depencency distance

For ADD computation, the ChatGPT-o1 model performs closest to the gold standard with small differences between the Gemini and GPT-4o model.

[1] From `https://www.diva-portal.org` and `https://www.mimersbrunn.se/`, respectively.

[2] `https://stanfordnlp.github.io/stanza/`

Model	DD gold standard (per word)
Gemini	2.09
GPT-4o	2.11
ChatGPT-o1	1.67

Table 1: Dependency Distance Differences Between Models and gold standard

In addition, it should be noted that GPT-4o consistently ignored all punctuation marks (commas, full stops, and parentheses) in its generated dependency trees. Gemini was inconsistent, including punctuation in some cases, but skipping them in other cases. Only ChatGPT-o1 included all punctuation marks in all the trees, exactly as in the gold standard trees.

Model	MMLU Score	LIX Error
llama-11b	73.0	7.65
llama-1b	32.2	17.89
llama-3b	58.0	1.58
llama-90b	86.0	10.73
Gemini	85.9	8.40
GPT 4-o	88.7	16.10
ChatGPT-o1	92.3	1.90

Table 2: MMLU and LIX Scores for Various Models

5 Discussion

Our results reveal that while LLMs have made significant advancements in language generation, they still face challenges in computing certain metrics related to language complexity, particularly when these metrics require precise numerical operations or structural reasoning. This study explored two specific tasks: the computation of the LIX readability metric and Average Dependency Distance (ADD). Our findings demonstrate that the models show varying degrees of accuracy, with ChatGPT-o1-preview outperforming Gemini and GPT-4o in both tasks.

5.1 LIX Computation

ChatGPT-o1-preview was able to calculate the LIX score with the highest accuracy, achieving results very close to the gold standard. The discrepancies in Gemini and GPT-4o suggest that these models struggle more with basic arithmetic operations, such as counting the number of long words and sentences, and correctly applying the LIX formula. This points to limitations in their ability to handle tasks that require detailed token-level information, which can be lost in tokenization and the internal representation of text in LLMs.

The LIX task demonstrates that although LLMs have robust language generation capabilities, their numerical reasoning and token awareness can still be inconsistent. Tasks like LIX, which require explicit counting and classification, reveal that models can misinterpret or miscount tokens, likely due to the abstraction of text into subword tokens during model processing.

5.2 Dependency Parsing and ADD

For dependency parsing and ADD, all models demonstrated a reasonable capacity to parse sentences and calculate dependency distances. However, the average differences from the gold standard indicate that ChatGPT-o1-preview again produced the closest results, while both Gemini and GPT-4o had slightly larger deviations.

The complexity of dependency parsing involves recognizing syntactic structures and understanding relationships between words, which requires models to perform tasks typically associated with traditional NLP pipelines. Although LLMs are not explicitly designed for tasks like dependency parsing, their ability to handle it with reasonable accuracy highlights their evolving capabilities in structural reasoning. Nonetheless, the errors in dependency distance computation, even in the best-performing model, suggest that LLMs still lack the precision of specialized tools like Stanza or traditional dependency parsers.

5.3 Broader Implications

These findings have broader implications for the integration of LLMs into tasks that require both linguistic understanding and computational precision. While models like ChatGPT-o1-preview show promise, particularly in language-focused tasks, they still require further refinement to ensure their reliability in professional domains where exact calculations are essential, such as legal, medical, or educational applications.

Moreover, our results indicate that models trained for general language tasks may struggle with specific, structured tasks such as readability scoring or syntactic parsing, which require a deeper understanding of the internal structure of language. This suggests that LLMs may benefit from targeted fine-tuning or the incorporation of more explicit training data that focuses on such tasks.

5.4 Clinical Implications

This study is part of a larger research project aiming to link linguistic complexity to various medical conditions, such as Alzheimer's dementia and Amyotrophic lateral sclerosis (ALS). In clinical settings, language analysis can serve as an early diagnostic tool for cognitive decline. By understanding how language complexity correlates with certain diseases, clinicians can leverage language models to detect subtle changes in a patient's speech or writing, potentially flagging early signs of conditions like dementia.

LLMs have the potential to automate the analysis of language complexity in clinical narratives, patient interviews, or written texts, providing a non-invasive method for screening and monitoring patients over time. For instance, the LIX readability metric and Average Dependency Distance (ADD) could be used to track changes in a patient's linguistic abilities, offering insights into cognitive functions that might not be immediately apparent through standard clinical assessments. These methods could be added to chatbot interfaces or speech to speech interfaces powered to LLMs to automatically assess language ability.

However, our findings also highlight the current limitations of LLMs in performing precise linguistic computations, which is crucial for reliable clinical diagnostics. While models like ChatGPT-o1-preview show promise, further refinement is needed to ensure these tools are accurate and consistent enough for clinical use. Nevertheless, the potential integration of LLMs into clinical diagnostics represents a promising direction for future research, where these models could complement traditional assessment methods, providing clinicians with an additional layer of information to support early diagnosis and intervention.

5.5 Language complexity skill as a proxy for general ability

When evaluating language models, evaluating general ability is not straightforward and require either time intensive human evaluation or structured benchmarks that can be unreliable. Having a quick way to assess a models general ability would be useful both while training models and while evaluating them.

Our work show that more capable models perform better on language complexity evaluations. Language complexity ability measures language ability itself rather than measuring context. As such, language complexity skill could in theory be used as a noisy proxy for general ability. We hope to explore this further in future work.

5.6 Limitations and Future Work

One limitation of our study is that we only evaluated the models on Swedish texts, which may not generalize to other languages. Additionally, our focus on university and high school essays might not represent the full range of text complexities LLMs encounter in real-world applications. Future work could involve expanding the dataset to include more diverse text types, as well as incorporating other complexity measures, such as Flesch-Kincaid or Gunning Fog indexes. We also aim to do more analysis of the relationship between general ability of the models and language complexity ability.

Additionally, as LLMs evolve rapidly, ongoing assessments are necessary to determine how newer models compare in terms of accuracy and reliability on these types of tasks. Fine-tuning models specifically for linguistic analysis, or incorporating hybrid systems that combine LLMs with traditional rule-based systems for tasks like dependency parsing, may offer a path forward in improving both their precision and flexibility.

5.7 Conclusion

In conclusion, while ChatGPT-o1-preview demonstrated the best performance in both LIX computation and dependency parsing, all models exhibited certain limitations in handling language complexity tasks. These results underscore the need for continued improvements in LLMs to enhance their performance on tasks that combine language and numerical reasoning, and suggest that with further refinements, LLMs could be reliably employed in areas requiring precise linguistic analysis.

References

Carl-Hugo Björnsson. 1968. *Läsbarhet: hur skall man som författare nå fram till läsarna?* Bokförlaget Liber.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. In *Dependency parsing*, pages 11–20. Springer.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-wei Lin, and Yang Liu. 2024. Llms for relational reasoning: How far are we? *arXiv preprint arXiv:2401.09042*.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Chinmay Mittal, Krishna Kartik, Parag Singla, et al. 2024. Puzzlebench: Can llms solve challenging first-order combinatorial reasoning problems? *arXiv preprint arXiv:2402.02611*.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024a. Can llms reason with rules? logic scaffolding for stress-testing and improving llms. *arXiv preprint arXiv:2402.11442*.

Zhihu Wang, Shiwan Zhao, Yu Wang, Heyuan Huang, Jiaxin Shi, Sitao Xie, Zhixing Wang, Yubo Zhang, Hongyan Li, and Junchi Yan. 2024b. Re-task: Revisiting llm tasks from capability, skill, and knowledge perspectives. *arXiv preprint arXiv:2408.06904*.

A Appendix: Prompts used

1. Complexity Measurement (LIX)

Analyze the complexity of the following text with the following formula:
Calculate the LIX (Läsbarhetsindex) score for Swedish text.
LIX = A + B, where: A = number of words / number of sentences B = (number of long words * 100) / number of words
Long words are defined as words with more than 6 characters.
Text: <insert text>
Please provide the result in JSON format with the following structure: "score": <LIX score>, "explanation": "<explanation of the calculation>"

2. Average Dependency Distance (ADD)

I would like you to print the dependency parsing result for a given Swedish sentence. Print the result with one word on each row, on the following form:
1, Han, 2, 1
2, köper, 0, 0
3, en, 4, 1
4, bok, 2, 2
where the first number is the word index, the second column is the word itself, the third column is the index of the head word, and the last number is the dependency distance (i.e. the absolute difference between the index and the head word index). The root word should have head word=0 with a distance of 0. Finally, print the average of all the dependency distances in the sentence. Here is the sentence: "*text*"

B Appendix: Texts used for the LIX computations

u1

Finns det någon mening med att studera historia?Jag anser att det gör det. Som ett argument för det kan man använda det klassiska uttrycket: - Man lär sig av sina misstag. Och det gör man. Misstag som man gjort ligger bakom en och det som ligger bakom en är också historia. ven om det rör sig om att man som 2-ring lär sig att inte springa in ett träd för att man får väldigt ont då.

u2

En dator brukar delas in i tre olika delar. Nämn dessa delar och förklara varför man valt just denna indelning. Datorn brukar som sagt delas in i tre olika delar. En centralenhet, en indataenhet och en utdataenhet. Indataenheter är som det låter, saker som vi använder för att skicka in data till datorn, tangentbordet är ett bra exempel, scanner och gamepads är två andra exempel.

u3

Pizzans historia börjar antagligen såhär:"Det var mycket strider förr. Folk reste omkring, men tallrikarna blev smutsiga, och diska dem tog för lång tid. Men så var det nån smart hjärna som kom på att man kunde göra tallrikarna av bröd! Ja, dom åt på tallrikarna av bröd, men efter måltiden slängde dom brödet. Långt senare började en resturang,

kallad som Bruno vid berget Vezuvio, att lägga tomater, mozarella och persilja på, eller om det var nån annan krydda. Men den var endast känd där, i byn.

u4

Historien om WUFC handlar om ett av de största graffiticrewen i Stockholm. Journalisten Björn Almqvist har följt några av graffitimålarna som är med i WUFC under flera års tid och fotat alla tunnelbanor som de har målat på, alla väggar de har målat på och på deras resor runt om i världen. Min personbeskrivning på graffitimålaren QUE skulle vara att han håller på med graffiti för att uttrycka sig själv och sina känslor. Que är inte den som tar till våld i första taget utan försöker att lösa sina konflikter med ord. Det är lite svårt att göra en miljöskrivning ur denna bok, men om jag skulle vilja beskriva någon plats skulle det vara hemma hos Que.

u5

Jag valde att skriva om när vallonerna emigrerade hit från Vallonien till Sverige. Dels valde jag det för att de har spelat en stor roll i vår svenska smidesindustri, dels för att jag fick höra att vi hade vallonblod i släkten och tyckte därför att det var ett intressant ämne.

Vallonerna emigrerade hit från Vallonien, ett område som ligger mellan Belgien och Frankrike, under 1600-talet och en bit framåt. När jag gick över statistik om emigration från Vallonien fann jag att Sverige var bland de fem regioner dit vallonerna flyttat till mest, tillsammans med Flandern, Brasilien, Argentina och USA (Wisconsin, framförallt). Man kan då fråga sig vad orsakerna var till att Sverige var så lockande för valloner och varför de emigrerade just hit.

h1

Under 1800-talets slut bodde svenskarna på landet och det var bara två av tio som bodde i städerna. Vid år 2000 bodde det så mycket som nio av tio av svenskarna i städer och tätorter. Urbaniseringen, som detta kallas, har gjort så att hälften av dagens befolkning bor i de femton största städerna. Urbaniseringen har gjort många förändringar när det gäller bosättningen i landet under dessa år. Just nu bor cirka 85 % av befolkningen i städerna och i tätorter.

h2

FN bildades den 24 oktober år 1945. Deras föregångare var Nationernas förbund, men dom lyckades inte så bra. Nationernas förbund klarade inte va pressen efter första världskriget. Efter 2: a världskriget så bestämde sig 51 länder att bilda FN, Förenta nationerna. I början av 1994 så var de 184 länder som var medlemmar i FN.

h3

I Nigeria varierar klimatet väldigt mycket beroende på var i landet man befinner sig. I söder alltså där Nigeria har kontakt med Atlanten är det varmt året om, ungefär 25 grader. Där regnar det också väldigt mycket ungefär 2000-3000mm per år. Anledningen att det regnar så mycket där är att Nigerias syd kust ligger så nära Atlanten. På morgonen och lite in på dagen avdunstar vattnet ifrån Atlanten som senare regnar ned på eftermiddagen.

h4

Jag skulle vilja säga att imperialismen började så långt tillbaks som för2000år sedan, och då tänker jag främst på rommarna som hade erövrat stora delar Europa och även delar av Asien. Rommarna styrde dem kända världen som dem mest överlägsnaste ledaren i världen. Jag skulle även vilja kalla spanjorerna och portugiserna för imperialister. De tog över helavärldsdelar och tog dem som kolonier. Dessa länder ihop med tyskland ochEngland koloniserade även hela Afrika och i och med det kom också handeln med slavar som skeppades i massor till "den nya världen" Amerika.

h5

I Sverige har vi en av världens bästa lagar mot diskriminering inom arbetslivet. Det skall inte spela någon roll varifrån man kommer eller vilken hudfärg man har, det skall vara den mest kvalificerade på jobbet. Allting låter bra så här långt. Hur kommer det sig då att Sverige har ett yrke som består av tandläkare, läkare, ingenjörer och andra högutbildade? Det är inte någon ny specialutbildning för medicinstuderande utan taxichaufförer. Hur kommer detta sig?SituationMånga av de invandrare som kommer till Sverige har inte varit några outbildade bidragstagare i sitt hemland.

C Appendix: Sentences used for dependency parsing

1. Även om det rör sig om att man som 2-åring lär sig att inte springa in ett träd för att man får väldigt ont då.

2. Indataenheter är som det låter, saker som vi använder för att skicka in data till datorn, tangentbordet är ett bra exempel, scanner och gamepads är två andra exempel.

3. Långt senare började en resturang, kallad som Bruno vid berget Vezuvio, att lägga tomater, mozarella och persilja på, eller om det var nån annan krydda.

4. Journalisten Björn Almqvist har följt några av graffitimålarna som är med i WUFC under flera års tid och fotat alla tunnelbanor som de har målat på, alla väggar de har målat på och på deras resor runt om i världen.

5. När jag gick över statistik om emigration från Vallonien fann jag att Sverige var bland de fem regioner dit vallonerna flyttat till mest, tillsammans med Flandern, Brasilien, Argentina och USA (Wisconsin, framförallt).

6. Urbaniseringen, som detta kallas, har gjort så att hälften av dagens befolkning bor i de femton största städerna.

7. Efter 2: a världskriget så bestämde sig 51 länder att bilda FN, Förenta nationerna.

8. På morgonen och lite in på dagen avdunstar vattnet ifrån Atlanten som senare regnar ned på eftermiddagen.

9. Jag skulle vilja säga att imperialismen började så långt tillbaks som för2000år sedan, och då tänker jag främst på rommarna som hade erövrat stora delar Europa och även delar av Asien.

10. Många av de invandrare som kommer till Sverige har inte varit några outbildade bidragstagare i sitt hemland.