Wordnet and Word Ladders: Climbing the abstraction taxonomy with LLMs

Giovanni Puccetti[1], **Andrea Esuli**[1], **Marianna Bolognesi**[2]

[1]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
[2]ABSTRACTION Research Group – Università di Bologna,

Correspondence: giovanni.puccetti@isti.cnr.it

Abstract

WordNet has long served as a benchmark for approximating the mechanisms of semantic categorization in the human mind, particularly through its hierarchical structure of word synsets, most notably the IS-A relation. However, these semantic relations have traditionally been curated manually by expert lexicographers, relying on external resources like dictionaries and corpora. In this paper, we explore whether large language models (LLMs) can be leveraged to approximate these hierarchical semantic relations, potentially offering a scalable and more dynamic alternative for maintaining and updating the WordNet taxonomy.

This investigation addresses the feasibility and implications of automating this process with LLMs by testing a set of prompts encoding different sociodemographic traits and finds that adding age and job information to the prompt affects the model ability to generate text in agreement with hierarchical semantic relations while gender does not have a statistically significant impact.

1 Introduction

The advent of large language models (LLMs) has revolutionized the landscape of Natural Language Processing (NLP), providing new avenues for exploring linguistic structures and semantic relations. One area of interest is the hierarchical organization of word meanings, captured by the semantic relation IS-A represented in WordNet. This paper aims to investigate the capacity of LLMs to understand this semantic relation by inspecting their ability to construct word *ladders* based on it. A word ladder is a sequence of words ordered by hypernym/hyponym relation, that include an initial given word and that go from a more generic term to a more specific one, as shown in Figure 1. In this perspective, word ladders represent the "branches" of WordNet, spanning from highly specific words

Token: *parallelogram*
Abstraction Ladder: thing, object, shape, polygon, quadrilateral, *parallelogram*, rectangle, rhombus, square

(a) **Parallelogram**

Token: *creationism*
Abstraction Ladder: idea, theory, belief, philosophy, worldview, *creationism*, theism, monotheism, biblical, fundamentalist, young-earth

(b) **Creationism**

Figure 1: Examples of LLM-generated ladders for a concrete concept (a): **parallelogram** and an abstract one (b): **creationism**.

(e.g., "chihuahua") to highly general ones (e.g., "living creature"). By analyzing how LLMs construct these hierarchies of hypernyms/hyponyms, we explore mechanisms that govern conceptual categorization and sense-making processes for different types of words, namely: concrete and abstract ones. Additionally, we will explore the sensitivity of the word ladders produced by LLMs to sociolinguistic factors, manipulating the sociodemographic "profile" that the LLM is prompted to play.

Our study explores the following key questions:

1. What type of categorizations do LLMs rely upon, when generating word ladders?
2. Can they organize hypernym/hyponym semantic relations for concrete as well as for abstract concepts?
3. Can LLMs approximate different types of

Category	Variants
job	linguist, researcher, teacher, poet, writer
age	8, 12, 15, 18, 22, 25, 30, 40, 50, 70
gender	not specified, male, female

Table 1: Sociodemographic variants encoded in different system prompts.

speakers, hence generating different types of word ladders? What type of speaker better approximates the categorizations encoded in WordNet?

Overall, through this analysis, we aim to contribute to a deeper understanding of the interaction between linguistic structures and model behavior, shedding light on the implications for both NLP applications and theories of human cognition.

2 Theoretical Background

Word taxonomies, such as WordNet (Miller, 1995), provide a structured representation of the paradigmatic relationships between words, labelling semantic relations like hypernymy (generalization) and hyponymy (specialization). These relations in turn shed light on the conceptual mechanisms of conceptual categorization, a core property of human cognition (Murphy, 2024), which is facilitated by language (Rissman and Lupyan, 2023). The construction of word ladders, which depict the progression from general to more specific terms, as shown in Figure 1, is a valuable task for assessing the semantic (paradigmatic) competence of large language models (LLMs). This approach allows us to evaluate their ability to abstract and generalize across different levels of word meaning.

As a matter of fact, LLMs in recent years have demonstrated remarkable abilities in natural language generation, based on these models' incredible accuracy in predicting and adjusting predictions on upcoming words in context, therefore on a syntagmatic level. Their architecture enables them to produce contextually appropriate responses in various domains (Brown et al., 2020), nevertheless crucial differences with human performance persist. Recent studies have specifically focused on the ability of LLMs to perform semantic categorizations and abstractions. For instance, (Samadarshi et al., 2024) examined the performance of state-of-the-art large language models (LLMs) against expert and novice human players in the New York Times Connections word game, a game in which players have to group words together to form semantically coherent ad-hoc categories. The authors found that even the top model, GPT-4o, can only fully solve 8% of the games. The results show that human players, especially experts, significantly outperform even the most advanced LLMs in tasks involving categorization and abstraction, which rely on paradigmatic relationships in the lexicon. In other words, while LLMs can typically generate coherent and cohesive text by inserting plausible words within syntagmatic contexts, their grasp of deeper paradigmatic, semantic relationships often falls short of aligning with established linguistic frameworks (Radford et al., 2019). In another recent example, Arora et al. (2023) highlight the limitations of LLMs in recognizing nuanced semantic distinctions, indicating that while these models can engage in categorization, their performance varies depending on the complexity of the task and the dataset used. To mitigate these limitations, Moskvoretskii et al. (2024) show that LLMs' understanding of semantic relations benefits from training on WordNet-like data.

ladder	Specificity	Position
thing	1.25	1
object	0.5	2
shape	1.25	3
polygon	1.5	4
quadrilateral	1.75	5
parallelogram	2.0	6
rectangle	2.25	7
rhombus	2.25	8
rhomboid	2.2	9
	Quality	0.89

Table 2: Example ladder with specificity scores calculated for each word. The Quality is measured as the Pearson correlation coefficient between the specificity and the position columns.

While several works stress that the difference with humans is significant, there are clues that, through training on in-domain data, LLMs can understand taxonomy-like relations (Moskvoretskii et al., 2024).

Constructing word ladders of hypernyms and hyponyms presents distinct challenges when dealing with concrete versus abstract concepts. Concrete concepts, such as "banana," generally exhibit

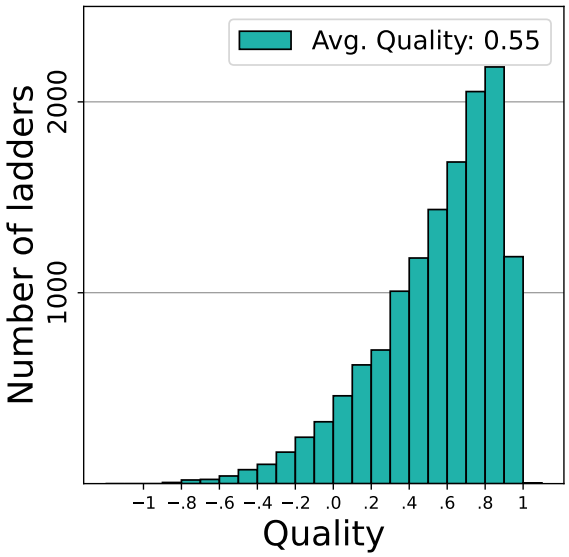


Figure 2: Distribution of the ladders' Quality for the *expert linguist prompt*.

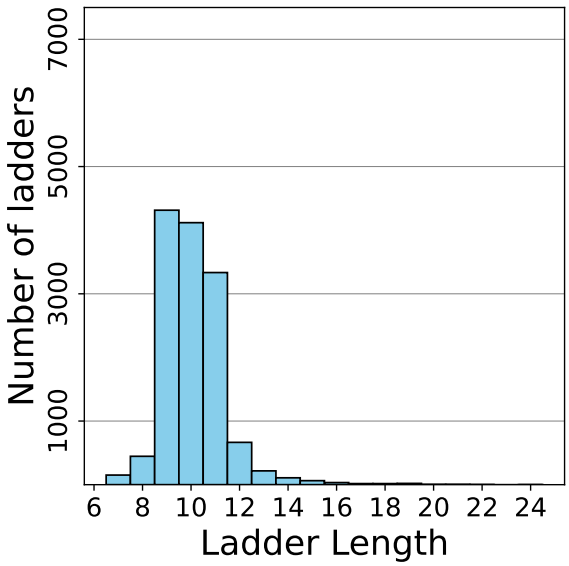


Figure 3: Distribution of the ladders' length for the *expert linguist prompt*.

clearer hierarchical relationships (Murphy, 2004; Mervis and Rosch, 1981), especially in domains like plants and animals, which are often structured by Linnaean taxonomies. For example, a "banana" is readily classified as a "fruit," which belongs to the broader category of "plant," and ultimately "living organism". In contrast, abstract concepts, like "belief", are more difficult to categorize due to their less tangible nature and variability in interpretation across contexts (Borghi et al., 2017b). Abstract concepts often involve multifaceted meanings that depend on cultural, social, and cognitive factors, making it harder to construct clear hierarchical relationships (Barsalou, 2008). This complexity can lead to ambiguity in determining appropriate hypernyms or identifying precise subcategories, as the boundaries between abstract concepts are more fluid than for concrete entities (Borghi et al., 2017a).

Finally, from a sociolinguistic perspective, research shows that sociodemographic factors significantly influence the types of categorizations performed by speakers, when using language (Labov, 1964; Milroy and Milroy, 1992; Barbieri, 2008; Wieling et al., 2011; Holmes, 2013). In the computational domain it has been shown that including demographic information such as age and gender significantly enhances the performance of text-classification tasks across multiple languages (Hovy, 2015). Ideally, by imposing sociodemographic profiles on LLMs, we can investigate how these factors influence the construction of word ladders and the resulting semantic relationships. Furthermore, we can correlate the specificity of words extracted from the generated ladders with that from WordNet, to infer which sociodemographic profiles better approximate the IS-A semantic relations encoded in WordNet. This approach not only enhances our understanding of LLM behavior but also aids in comprehending the peculiarities and potential limitations of WordNet, which is often used as a benchmark for evaluating various tasks, assuming that its cognitive underpinnings make it a suitable comparison to approximate any type of speaker (Bolognesi et al., 2020, inter alia).

To summarize, this study aims to systematically analyze the paradigmatic, hypernym/hyponym semantic relations encoded in the word ladders generated by LLMs. We will assess the accuracy and reliability of the categorizations produced, identify common challenges faced by the models, and explore how variations in sociodemographic profiles influence the semantic output.

3 Method

To explore the ability of LLMs to generate meaningful ladders in an open and replicable manner, we focus on *Llama 3.1 405b*, a highly performing open source LLM, which competes with proprietary models such as ChatGPT in several tasks.[1] A comparative study involving different models will be reserved for future research.

[1] https://huggingface.co/open_llm_leaderboard

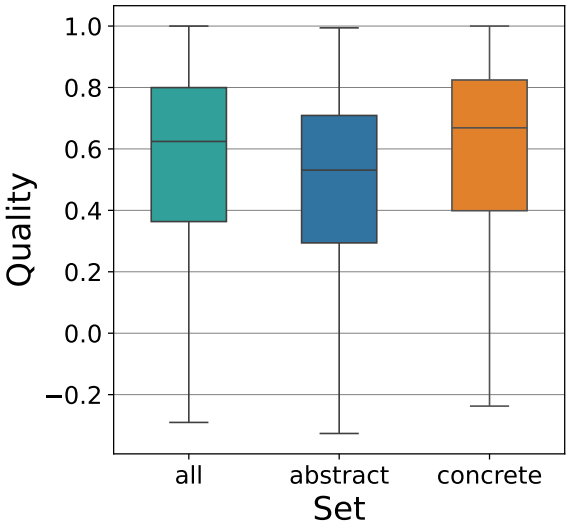


Figure 4: Boxplot of ladders Quality for the *expert linguist prompt*, comparing distribution when considering the full set (left), abstract nouns (center) and concrete nouns (right).

3.1 Ladders Generation

LLMs use two different prompts when generating text, a *system prompt* and a regular *prompt* (Dubey et al., 2024). The system prompt sets the overall behavior of the model, allowing it to adopt specific roles or tones. For instance, it might be instructed to act as "a lawyer specializing in maritime law" or "a cheerful person who frequently uses emoticons." This shapes the model's general response style but doesn't specify the particular task it should perform. The regular prompt, on the other hand, defines the specific task we want the model to execute. In our approach, we use the system prompt to simulate different sociodemographic profiles, while the regular prompt will guide the model to generate content aligned with our specific interests.

We instruct *Llama 3.1 405b* to generate ladders by using the following prompt: *Construct a list of single word concepts around the word: {word}. The bullets before the word {word} have to be increasingly more generic while those after the word {word} increasingly more specific. Make it look like one list.* In the instruction, *{word}* is replaced every time by a specific word. As the words for starting the ladder generation we use a list of 13,518 tokens, which are the items classified as nouns in the dataset of concreteness ratings collected by Brysbaert et al. (2014).

We explore two dimensions in sociodemographic profiles, age and job. We define 10 age values and 5 jobs (Table 1, see Appendix A for the complete list of all the system prompts), producing 15 system prompts: *You are a teenager of 18 years old learning in college*, *You are an expert linguist analysing the abstraction and concreteness of words*. We also generate 30 additional system prompts with an explicit specification of gender (male or female): *You are a young woman of 22 years old learning in university.* As a result we generate 13,518 $\times$ 45 = 608,310 ladders.[2]

3.2 Ladders Evaluation

We evaluate the ladders generated by LLM by calculating the Specificity of each word inserted in a ladder and correlating this measure with the order they have in the ladder. We use the measure of word specificity from Bolognesi et al. (2020). This metric is based on WordNet 3.0, which is available in the Natural Language Toolkit (NLTK, version 3.2.2) Python library (Bird et al., 2009). The measure is based on the distance of a word from the root node of the WordNet hierarchy, where the root is the most general concept, i.e., *entity*:

$$\text{Specificity}(w) = \frac{1+d}{20}$$

where d is the number of nodes between the word w and the root node, and 20 is the longest distance between the root node and a leaf in WordNet.

We define the quality of a ladder as how much the order of words in the ladder correlates with their order by Specificity measured using WordNet. Formally, if $W = \{w_i\}_{i=1}^{n}$ is a ladder composed of n words w_i and $X = \{x_i\}_{i=1}^{n}$ are their Specificity scores measured in WordNet, e.g. $x_i = \text{Specificity}(w_i)$, we define the quality of the ladder as:

$$\text{Quality}(X) = personr(X, N)$$

where $N = \{1, \ldots, n\}$ are the integers between 1 and n.

This *Quality* metric goes from -1 to 1, and it assigns scores close to 1 to the ladders where the words are sorted according to the Specificity measured in WordNet, 0 to those that ordered randomly and -1 to those that have a reverse order compared to their Specificity.

Table 2 shows an example of a ladder with the Specificity scores for each word and the Quality of

[2]We release this dataset and the code used to create it in anonymized form here https://anonymous.4open.science/r/abstract_llm-6B9A/README.md.

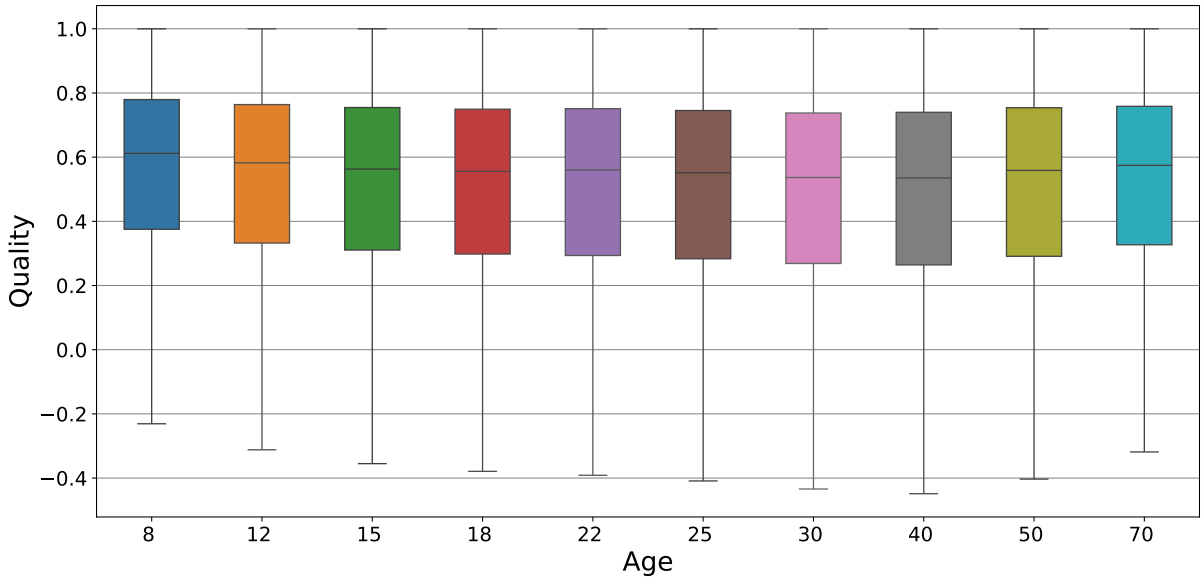


Figure 5: Boxplot of ladders Quality for all ages studied.

the ladder. In the example, the words are almost perfectly sorted and the Quality is high, however, there are two exceptions, *object*, which has a lower Specificity than *thing* and should therefore come first and *rhomboid* that has a lower Specificity than *rectangle* and *rhombus* and should therefore come before them, thus Quality is lower than 1.

4 Results

Before exploring separate sociodemographic roles, we want to understand how well the model is able to generate coherent and well ordered ladders, to do this, we initially focus on the ladders generated by the model when prompted as an expert linguist.

Figure 2 shows the distribution of Quality for all the ladders generated by the linguist prompts, there is a noticeable correlation between the Specificity of the entities in the ladder and their position, showing that they are well sorted. Notice how the average of 0.55 is relatively high in $[-1, 1]$ range, indicating a good degree of correlation. There are only a few ladders that have a negative Quality index, underlining that the model is rarely misaligned with WordNet, and therefore general commonsense.

Looking at the distribution of the lengths of the generated ladders, Figure 3 shows that the model spontaneously generates ladders mostly of length 9, 10, 11 although we don't ask for this explicitly in the prompt.[3]

[3]Indeed, the examples we provide in Figure 1 have length 9 and 11.

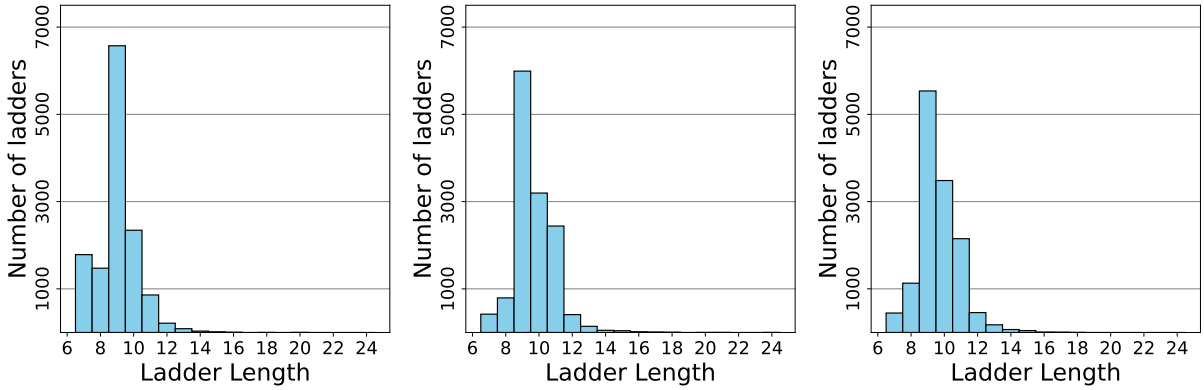
Abstractness and Concreteness: can the model generate ladders for both, abstract as well as concrete words? To answer this question we measure the Quality of the ladders on two subsets of the Brysbaert nouns, one containing more concrete examples and one containing more abstract ones, both built using WordNet (Bolognesi et al., 2020). Specifically, the former contains all the synsets that have the node "physical entity" as an ancestor, and the latter contains all the synsets that have the node "abstraction" as an ancestor.

Figure 4 shows the box plots for ladders Quality for the full set (green), only abstract nouns (blue), and only concrete nouns (orange). When compared to the full set of nouns, the Quality is higher for the concrete nouns, which have higher median, and lower for abstract nouns. This is coherent with human behaviour (Mervis and Rosch, 1981).

We conducted Mann-Whitney U tests to reject the null hypothesis that the medians of the concrete and abstract groups are the same. The tests returned very low p-values (of the order of 10^{-80}). This finding suggests that the difference in Quality among ladders constructed starting from abstract and concrete prompts is reflected in the LLMs. This is comparable to human behavior, where humans display more difficulties in creating taxonomic relations for abstract vs. concrete concepts.

5 A Multifaceted Perspective

WordNet was developed by a team of experts in linguistics, cognitive science, and lexicography.



(a) Child: 8 years old. (b) Adult: 30 years old. (c) Senior: 70 years old.

Figure 6: Distribution of ladders length for all nouns, for three different system prompts, (a) for a 8 years old child, (b) for a 30 years old adult, and (c) for a 70 years old senior.

Accordingly, our previous analyses prompted the LLM to mimic this type of speaker. We now extend our investigation to explore how the results change when the LLM is used to replicate the behavior of different sociodemographic groups.

In the analyses hereby reported we manipulate three sociodemographic factors, namely: Age, Profession, and Gender. We observe how these variables impact ladders construction.

Age: how does the age encoded in the system prompt affect the generated ladders? When prompting the model with different ages, we experimented with a wide range of ages: 8 years old, 12, 15, 18, 22, 25, 30, 40, 50, and 70. We used a finer categorization for younger ages and a coarser one for older ages, based on the assumption that during developmental and schooling years, more noticeable changes in language use occur compared to adulthood.

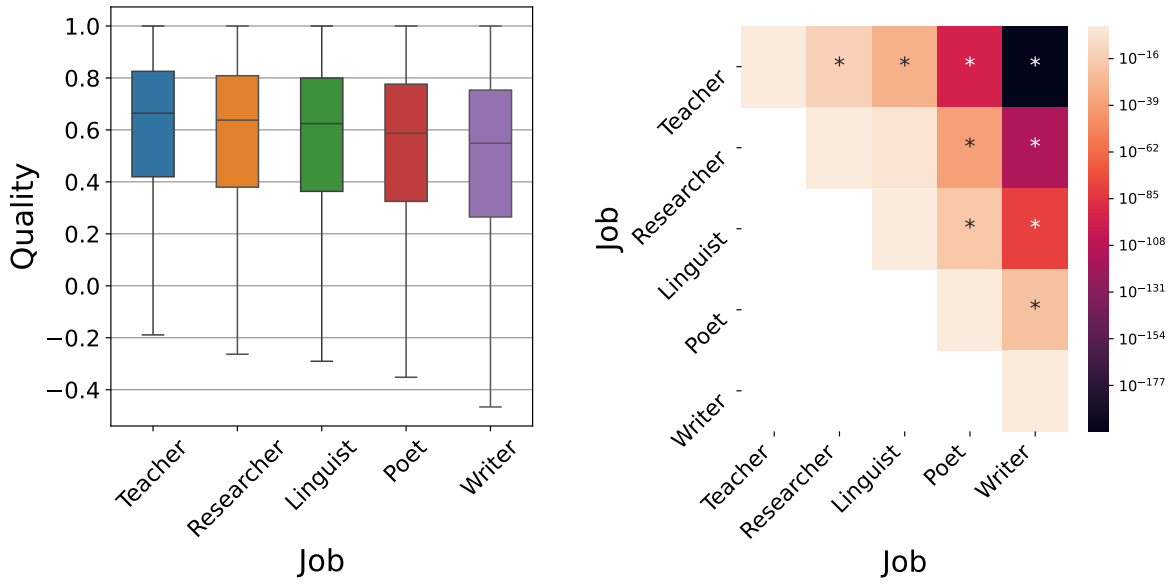
Figure 5 shows the distribution of quality of the ladders generated by the model when prompted to act like a person of different ages. We can see a *U-shape*, where the Quality appears to be higher, i.e., better correlated with WordNet-based Specificity, for young and old ages, while lower for ages between 22 and 50. We can thus conclude that the model is more aligned to WordNet when prompted as a child or as an older person.

This finding is somewhat counterintuitive, and we hypothesized that the model would generate shorter ladders when prompted with a "child" or "senior" profile, based on the idea that both these groups might find it harder to generate longer ladders. Figure 6 shows the distribution of ladders length (number of words inserted in a ladder) for system prompts where the age is set to 8 years old (fig. 6a), 30 years old (fig. 6b) and 70 years old (fig. 6c). The model tends to generate ladders of length 9, 10 and 11, for all ages. Note that we did not specify the length in the prompt, the model spontaneously keeps the lengths in this range. However, when prompted to mimic the behavior of a child or senior, ladders appear to be shorter than when prompted to mimic the behavior of an adult. To ensure that we are not seeing a spurious difference among different ages, we performed a one-way ANOVA test on the ladders length for the different ages, the test returned a p-value of the order 10^{-44} indicating that the difference is significant and the post-hoc Tukey test also returns only significant p-values with the highest of the order of 10^{-10}.

Profession: how strongly does encoding the job in the system prompt affect ladders generation? To investigate the effect of different expertise, we compare prompts that ask to act like a person with a specific job. We use the following jobs, all involving intellectual work: *linguist*, *poet*, *teacher*, *writer*, *researcher*.

Figure 7 shows the box plots of the ladders Quality for each job. Interestingly, the two roles involving more creative work (poet and writer) are the ones with the lower Quality, while the more analytical roles, i.e. linguist, researcher and teacher, have higher Quality. These results are coherent with WordNet systematicity, favoring more analytical writing, and also with the inherent fuzziness of the concept of specificity which becomes harder to understand when used in more creative writing.

To test the significance of the differences among



(a) Boxplot of ladders concreteness for all jobs studied.

(b) P-values for pairwise MannWithney U tests.

Figure 7: Study of the Quality distribution for all the jobs studied. In (a) the boxplot of the Quality for each job and in (b) a heatmap reporting the p-values for pairwise Mann-Whitney U tests (the * indicates p-values lower than 0.05).

the distributions, we performed pairwise Mann-Whitney U tests (McKnight and Najab, 2010) among the ladders of each pair of prompts. This test determines whether two distributions are the same or not. In our case, when given two sets of ladders, we aim to answer the question: "Do these two sets of ladders reflect an equal ability to understand the concepts of specificity?" Performing the Mann-Whitney U test between the characteristics of two sets of ladders generated from different system prompts allows us to address this question. We also apply the Bonferroni correction (Dunn, 1961) to adjust for multiple comparisons.

Through this approach we want to understand if prompting the model to behave according to different sociodemographic classes generates significantly different ladders and what type of sociodemographic profile approximates the word specificity encoded in WordNet.

Figure 7b shows the p-values for the Mann-Whitney U tests. The only non-significant p-value (above 0.05 after applying Bonferroni correction) relates to the comparison between Linguist and Researcher, which are interestingly very close types of profession, with the profile "Teacher" being associated with the highest ladders Quality.

Gender: how strongly does encoding gender in the system prompt affect ladders generation? To understand how gender affects the Quality of the generated ladders, we had the model generate responses using the same prompts as for age, with an added description of the character as either female or male. For example, we used “boy/girl” for younger ages, “man/woman” for middle ages, and “male/female” for older ages. We compared all the ladders generated across all ages. Similarly to how we compared different jobs, we used a Mann-Whitney U test with Bonferroni correction to determine whether adding different genders to the system prompts affects the generated ladders.

Figure 8 shows the pairwise p-values for the Mann-Whitney U tests of the null hypothesis that the distribution of qualities is the same between pairs of system prompts. The squares marked with an asterisk indicate that the null hypothesis is rejected, meaning there is a significant difference between the distributions of the two groups. While this is true for most comparisons, there are interesting exceptions.

Most notably, male and female ladders generated at the same ages are not significantly different from each other, this is shown in the upper diagonal starting at the intersection between the rows "8 female" and "8 male" and marking all intersections between prompts with the same age, showing that gender alone does not change the overall ability of the model to generate coherent ladders. [4]

[4]We made the same test also for the specification of jobs

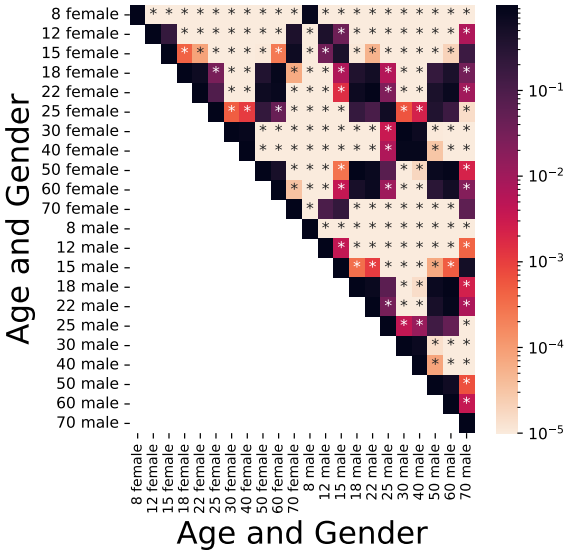


Figure 8: P-values for the Mann Withney U tests to measure if the distribution of qualities is significantly different across prompts, the * indicates p-values below 0.05 (The heatmap uses log-scale for more understandable coloring, while * are based on the actual p-values).

Figure 8 also reveals that the "U - shape" shown in Figure 5, indicating how the higher Quality is seen for younger and older ages, appears even if we are diversifying by gender. Indeed, we see that outside the upper diagonal the only non significant p-values are of the kind younger female or male against older female or male. Indicating a degree of similarity between the qualities of younger and older characters.

Note that the high number of significant tests (even when using Bonferroni correction) is reasonable given the large sample size of the populations (of ladders) we are comparing, returning small p-values also with small distribution shifts.

6 Discussion and Conclusions

This study is focused on the ability of Large Language Models (LLMs) to replicate the hierarchical structure of word meaning representations found in WordNet and summarized in the semantic relation IS-A. This relation connects more specific terms (hyponyms) to more general categories (hypernyms). We explored to what extent do LLMs are able to reproduce this paradigmatic semantic relation, and in particular their ability to do so for concrete and for abstract concepts. Moreover, we explored how different sociodemographic profiles prompted to the LLM approximate the categorizations encoded in WordNet.

We can summarize the main results as follows:

- LLMs overall replicate humans' difficulties in constructing word taxonomies for abstract concepts compared to concrete ones.
- When prompted to impersonate different jobs LLM generate significantly different ladders resulting in varying Quality. More analytical jobs generate ladders that are more aligned with WordNet, while more creative roles generate ladders with lower Quality;
- Age plays a relevant role when prompting LLMs to use their understanding of specific and generic concepts and we identify a "U - pattern" where younger and older ages result in higher Quality, we speculate this is due to simpler ladders that are more easily aligned with WordNet architecture;
- Gender does not play a major role in the generation of ladders, since adding male or female attributes to the prompts doesn't significantly affect the Quality of ladders.

We also acknowledge the main limitations of this study: 1. While we compared the ladders generated by the LLM to those from WordNet, we are unable to make direct comparisons with human-generated ladders due to the absence of such data; 2. Although we tested several system prompts, there is potential for further exploration with more complex and diverse sociodemographic profiles. We plan to address all these points in future research.

In conclusion, this study sheds light on our understanding of both the capabilities and limitations of Large Language Models (LLMs) in categorizing and abstracting knowledge based on semantic lexical relations. By analyzing the ability of LLMs to generate word ladders that mirror WordNet's IS-A hierarchies, the research helps us understand how these models handle complex semantic structures.

The findings will not only help identify where LLMs excel or fall short in replicating human-like categorization but also highlight the nuanced challenges they face, particularly in distinguishing between concrete and abstract concepts.

and we obtained the same result, i.e., gender specification has no significant impact in the characteristics of the generated ladders.

Acknowledgments

Financed by the European Union - NextGenerationEU through the Italian Ministry of University and Research under PNRR - PRIN 2022 (2022EPTPJ9) "WEMB: Word Embeddings from Cognitive Linguistics to Language Engineering and back", and by the European Union (GRANT AGREEMENT: ERC-2021-STG-101039777). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

Federica Barbieri. 2008. Patterns of age-based linguistic variation in american english. *Journal of Sociolinguistics*, 12(1):58–88.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Beijing.

M. Bolognesi, C. Burgers, and T. Caselli. 2020. On abstraction: decoupling conceptual concreteness and categorical specificity. *Cognitive Processing*, 21(3):365–381.

Anna M. Borghi, Laura Barca, Ferdinand Binkofski, and Luca Tummolini. 2017a. Abstract concepts, language and sociality: From acquisition to inner speech. *Topics in Cognitive Science*, 9(3):673–693.

Anna M. Borghi, Ferdinand Binkofski, Cristiano Castelfranchi, Felice Cimatti, Claudia Scorolli, and Luca Tummolini. 2017b. The challenge of abstract concepts. *Psychological Bulletin*, 143:263–292.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Marc Brysbaert, AB Warriner, and V Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *BEHAVIOR RESEARCH METHODS*, 46(3):904–911.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer

Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

Janet Holmes. 2013. *An Introduction to Sociolinguistics*. Routledge.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

William Labov. 1964. *The social stratification of English in New York City*. Ph.D. thesis, Columbia University. Ph.D. thesis.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

Carolyn B. Mervis and Eleanor Rosch. 1981. Categorization of natural objects. *Annual Review of Psychology*, 32:89–115.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Lesley Milroy and James Milroy. 1992. Social network and social class: Toward an integrated sociolinguistic model. *Language in Society*, 21(01):1–26.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.

Gregory L. Murphy. 2004. *The big book of concepts*. MIT Press.

Gregory L. Murphy. 2024. *Categories We Live By: How We Classify Everyone and Everything*. The MIT Press, Cambridge, MA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Preprint*, arXiv:1901.11117.

Lauren Rissman and Gary Lupyan. 2023. The power of the lexicon: Eliciting superordinate categories with and without labels. *PsyArXiv*. Preprint.

Prisha Samadarshi, Mariam Mustafa, Anushka Kulkarni, Raven Rothkopf, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Connecting the dots: Evaluating abstract reasoning capabilities of llms using the new york times connections word game. *arXiv preprint arXiv:2406.11012v2*. Cs.CL.

Martijn Wieling, John Nerbonne, and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLOS ONE*, 6(9):e23613.

A Prompts

Table 3 shows that full list of prompts used to generate the ladders.

Generic Prompts
You are a child of 3 years old learning about the world. You are a child of 5 years old learning about the world. You are a child of 8 years old learning in elementary school. You are a child of 12 years old learning in middle school. You are a teenager of 15 years old learning in high school. You are a teenager of 18 years old learning in college. You are a young adult of 22 years old learning in university. You are a young adult of 25 years old learning in graduate school. You are a young adult of 30 yearnewpages old learning in a professional setting. You are a middle-aged adult of 40 years old learning in a professional setting. You are a middle-aged adult of 50 years old working in a professional setting. You are a middle-aged adult of 60 years old working in a professional setting. You are a senior of 70 years old who is now retired.
You are a teacher explaining the concept of abstraction and concreteness to a class of 5th grade students. You are a researcher studying the concept of abstraction and concreteness in language. You are an expert linguist analysing the abstraction and concreteness of words.
You are a poet trying to find the perfect words to describe a feeling. You are a writer trying to find the perfect words to describe a scene.
Female Prompts
You are a girl of 3 years old learning about the world. You are a girl of 5 years old learning about the world. You are a girl of 8 years old learning in elementary school. You are a girl of 12 years old learning in middle school. You are a female teenager of 15 years old learning in high school. You are a female teenager of 18 years old learning in college. You are a young woman of 22 years old learning in university. You are a young woman of 25 years old learning in graduate school. You are a young woman of 30 years old learning in a professional setting. You are a middle-aged woman of 40 years old learning in a professional setting. You are a middle-aged woman of 50 years old working in a professional setting. You are a middle-aged woman of 60 years old working in a professional setting. You are a senior woman of 70 years old who is now retired.
You are a female teacher explaining the concept of abstraction and concreteness to a class of 5th grade students. You are a female researcher studying the concept of abstraction and concreteness in language. You are a female expert linguist analysing the abstraction and concreteness of words.

You are a female poet trying to find the perfect words to describe a feeling. You are a female writer trying to find the perfect words to describe a scene.
Male Prompts
You are a boy of 3 years old learning about the world. You are a boy of 5 years old learning about the world. You are a boy of 8 years old learning in elementary school. You are a boy of 12 years old learning in middle school. You are a male teenager of 15 years old learning in high school. You are a male teenager of 18 years old learning in college. You are a young man of 22 years old learning in university. You are a young man of 25 years old learning in graduate school. You are a young man of 30 years old learning in a professional setting. You are a middle-aged man of 40 years old learning in a professional setting. You are a middle-aged man of 50 years old working in a professional setting. You are a middle-aged man of 60 years old working in a professional setting. You are a senior man of 70 years old who is now retired.
You are a male teacher explaining the concept of abstraction and concreteness to a class of 5th grade students. You are a male researcher studying the concept of abstraction and concreteness in language. You are a male expert linguist analysing the abstraction and concreteness of words.
You are a male poet trying to find the perfect words to describe a feeling. You are a male writer trying to find the perfect words to describe a scene.

Table 3: Prompts used to generate the ladders.

B Ladders statistics for all ages

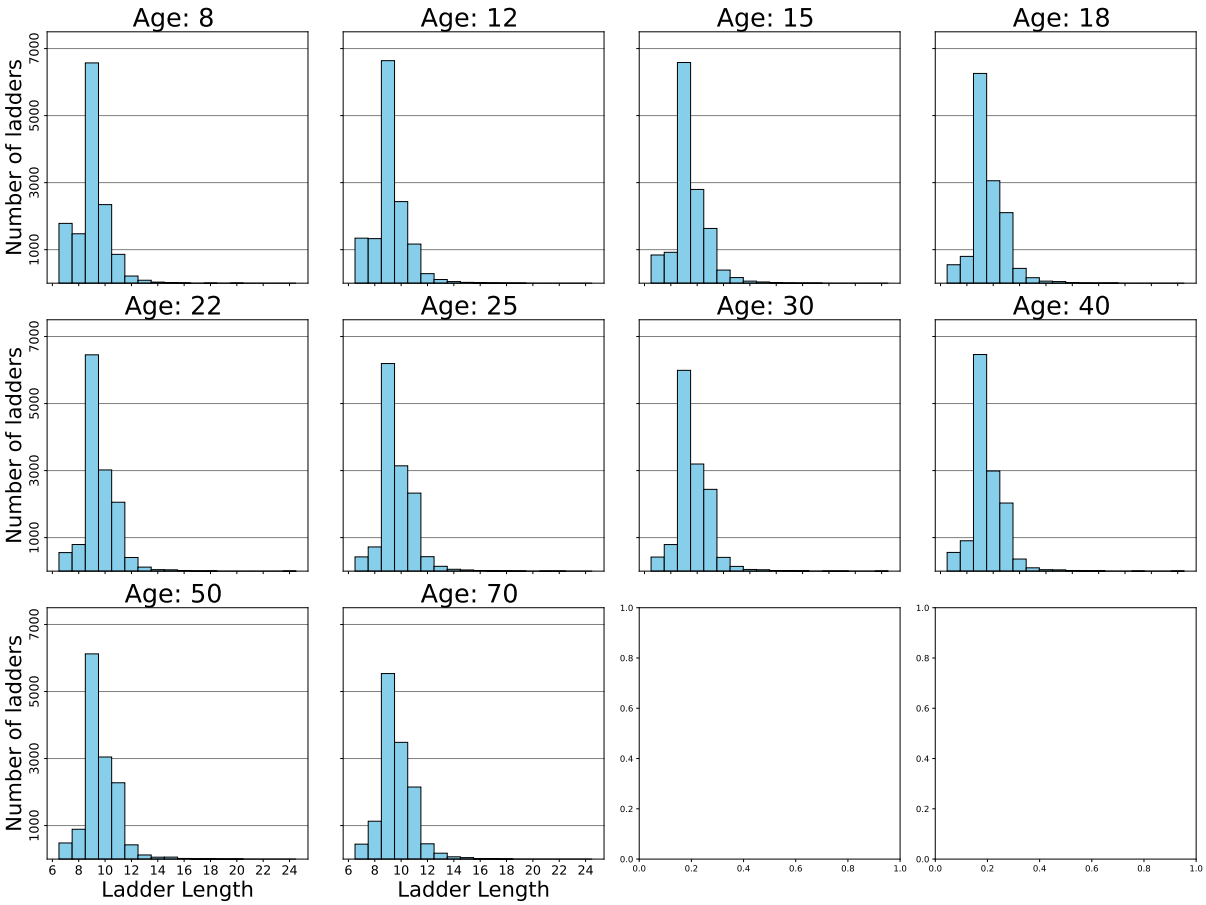


Figure 9: Ladder length distribution for all ages tested.