Early Findings in Using LLMs to Assess Semantic Relations Strength

André Fernandes dos Santos
CRACS & INESC Tec LA / Faculty of Sciences, University of Porto, Portugal

José Paulo Leal
CRACS & INESC Tec LA / Faculty of Sciences, University of Porto, Portugal

Abstract

Semantic measure (SM) algorithms allow software to mimic the human ability of assessing the strength of the semantic relations between elements such as concepts, entities, words, or sentences. SM algorithms are typically evaluated by comparison against gold standard datasets built by human annotators. These datasets are composed of pairs of elements and an averaged numeric rating. Building such datasets usually requires asking human annotators to assign a numeric value to their perception of the strength of the semantic relation between two elements. Large language models (LLMs) have recently been successfully used to perform tasks which previously required human intervention, such as text summarization, essay writing, image description, image synthesis, question answering, and so on. In this paper, we present ongoing research on LLMs capabilities for semantic relations assessment. We queried several LLMs to rate the relationship of pairs of elements from existing semantic measures evaluation datasets, and measured the correlation between the results from the LLMs and gold standard datasets. Furthermore, we performed additional experiments to evaluate which other factors can influence LLMs performance in this task. We present and discuss the results obtained so far.

2012 ACM Subject Classification Computing methodologies → Language resources

Keywords and phrases large language models, semantic measures, semantic datasets

Digital Object Identifier 10.4230/OASIcs.SLATE.2024.4

Category Short Paper

Supplementary Material *Software (Datasets)*: `https://github.com/andrefs/punuy-datasets`
archived at `swh:1:dir:46a61b4a793d8a3ff57fb519af5a2c1a328368d7`
Software (Source Code): `https://github.com/andrefs/punuy-eval`
archived at `swh:1:dir:24f38b617eb018c501c0c8e5fd4b564c8727ed5b`

Funding This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.
André Fernandes dos Santos: Ph. D. Grant SFRH/BD/129225/2017 from Fundação para a Ciência e Tecnologia (FCT), Portugal.

1 Introduction

Semantic measures (SMs) are metrics designed to compare entities according to their semantics, i.e. their meaning [12]. SMs can be applied to units of language (e.g. words, sentences, documents), concepts, or instances (as long as they are semantically categorized – diseases, genes, geographical locations), and provide machines with the ability to estimate the strength of the semantic relationship between semantic entities.

SMs are based on the analysis of *semantic proxies* from which *semantic evidence* can be extracted, which must be somehow related to the *meaning* of the entity. For example, words that appear frequently close to each other on a text corpus are more likely to be related that

13th Symposium on Languages, Applications and Technologies (SLATE 2024).
Editors: Mário Rodrigues, José Paulo Leal, and Filipe Portela; Article No. 4; pp. 4:1–4:9
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

words that do not. The same can be observed for nodes on a knowledge graph which have many and/or shorter connections between them.

Despite their undeniable subjective component, with personal and cultural background playing their part in influencing our similarity appreciations, inter-human agreement on semantic similarity rating is high – the literature reports levels of 73% to 89% [32].

The accuracy of SMs is usually measured through comparison with averaged values reported by humans. This comparison is done by calculating the correlation between the expected and the obtained values, which allows focusing on covariance rather than the absolute values. If only the rank of the predictions is relevant, usually Spearman's rank correlation ρ is used; otherwise, Pearson's linear correlation coefficient r. Many gold standard datasets, typically composed of pairs of elements and their average semantic relation score, have been compiled and published in the literature. The full list of datasets we used for this work, too large to fit in this paper, is available online[1].

Large Language Models (LLMs) are computational models, trained on large text corpora, designed for text generation. LLMs are usually built using deep learning algorithms, such as recurrent neural networks or transformer models. LLMs have gained significant attention in the more recent years due to their ability to perform with high accuracy and performance typical natural language processing tasks. The notoriety of LLMs started first with the evolution of OpenAI's GPT models, then with the release of ChatGPT [16, 19] and with the development of open source models and models from other providers [21]. LLMs have been used for such diverse tasks such as language translation, question answering, summarization and text generation [11].

In this paper we present preliminary work we performed while attempting to answer the question: Can LLMs be used to accurately predict semantic measure values? To answer this we first tested how sensitive LLMs were to prompts variations (Section 2.1). Then we measured the correlation of LLMs SM predictions with multiple human-generated annotations (Section 2.2). Finally we tested whether LLMs had prior knowledge of the datasets and if this knowledge could be used (Section 2.3). Section 3 analyzes the results obtained and provides additional details about ongoing work.

2 LLM probing

Quick anecdotal experiments suggested that LLMs can predict semantic distance between concepts. For example, when we asked ChatGPT [34] to attribute a numeric value between 1 and 10 to the semantic relatedness of the concepts *computer* and *keyboard*, it answered "8", a value which intuitively seems adequate.

In this section we describe several structured approaches for probing LLMs to understand the extend of their capabilities in predicting semantic measures. These were implemented using the model providers APIs, taking advantage of the *function calling* feature [26, 2, 23] to obtain structured results. The model providers included were OpenAI [27] (`gpt-3.5-turbo`, `gpt-4`, `gpt-4-turbo`), Mistral [22] (`mistral-large`, `open-mixtral-8x22b`) and Anthropic [1] (`claude-3-sonnet`, `claude-3-opus`). The models were picked from the highest score models from publicly available leaderboards [5, 18]. Google's `gemini` models [9] and Cohere's `commandR` [6] were considered but eventually excluded due to limitations in their API.

[1] https://github.com/andrefs/punuy-datasets/tree/v6.1.2

Table 1 Average correlation between LLMs and datasets using different prompts.

English 0.71							Portuguese 0.64					
relatedness 0.63			similarity 0.76				relatedness 0.69			similarity 0.57		
P_{basic}	P_{scale}	P_{ws353}	P_{basic}	P_{scale}	P_{ws353}	P_{sl999}	P_{basic}	P_{scale}	P_{survey}	P_{basic}	P_{scale}	P_{survey}
0.52	0.68	0.71	0.55	0.79	0.84	0.90	0.67	0.61	0.80	0.40	0.64	0.64

2.1 Prompt variations

Prompt engineering refers to the methodologies and techniques for crafting the optimal prompt for a given task. For evaluating how sensitive an LLM would be to different prompts, we asked several LLMs (`gpt-4-turbo`, `claude-3-opus`, `mistral-large` and `open-mixtral-8x22b`) to rate the pairs of a few datasets, using different languages (Portuguese and English), semantic measures (similarity and relatedness), and prompt verbosity levels:

P_{basic} is very succinct and asks the model to rate on a scale from 1 (min) to 5 (max),

P_{scale} provides an explanation for each value on the scale (see Listing 1),

P_{ws353} was adapted from the WS353 [8] dataset, and includes a more detailed explanation of what the goal is,

P_{sl999} was adapted from the SimLex999 [14] dataset and includes examples of pairs of words that are very similar or dissimilar,

P_{survey} , developed by us to be used on a new Portuguese dataset we are currently building, includes examples of pairs of words and their expected rating.

Listing 1 Truncated example of a P_{scale} prompt.

```
Indicate how strongly the words in each pair are similar in meaning
using integers from 1 to 5, where the scale means: 1 - not at all
similar, 2 - vaguely similar, 3 - indirectly similar, 4 - strongly
similar, 5 - inseparably similar.
Pairs of words:
  - tap, knock
  - bruise, split
  - sell, market [...]
```

We repeated each query 10 times, then measured the correlation between the values obtained and the values from the dataset (using Pearson's correlation coefficient), and averaged them for each prompt, measure type and language.

Table 1 presents the averaged results. We can observe that generally, the longer and more detailed prompts obtain better results (P_{basic} gets the worst results; P_{scale} results are a little bit better and P_{ws353}, P_{sl999} and P_{survey} get the best results). The measure type seems to have little to no impact on results. The results for prompts and datasets in English are better than the ones in Portuguese. Due to constraints on the article size we could not include the individualized results for each model. We could however notice that the performance of the prompts did not always follow the global averages. For example, the `mistral-large` model, for the Mturk287 dataset [30] (English, relatedness) obtained a correlation of -0.14 with P_{basic}, 0.96 with P_{scale} and 0.34 with P_{ws353}.

Table 2 Confusion matrix for semantic similarity values for the MC30 subset.

LLM	gpt-3.5-turbo	gpt-4	gpt-4-turbo	claude-3-sonnet	claude-3-opus	mistral-large	open-mixtral-8x22b	MC30	RG65	WS353	PS65	LLM average
gpt-3.5-turbo		0.94	0.94	0.94	0.94	0.87	0.91	0.92	0.89	0.89	0.90	0.90
gpt-4			0.97	0.96	0.97	0.92	0.93	0.93	0.94	0.91	0.94	0.93
gpt-4-turbo				0.93	0.97	0.91	0.95	0.92	0.92	0.89	0.93	0.92
claude-3-sonnet					0.96	0.89	0.89	0.90	0.91	0.88	0.90	0.90
claude-3-opus						0.92	0.91	0.94	0.94	0.93	0.94	**0.94**
mistral-large							0.90	0.89	0.88	0.91	0.89	0.89
open-mixtral-8x22b								0.86	0.84	0.81	0.88	0.85
MC30									0.97	0.95	0.95	
RG65										0.91	0.96	
WS353											0.92	
PS65												

2.2 Comparing results for MC30

Most existing SM datasets are created and never replicated. Their authors publish them and frequently provide metrics such as the inter-annotator agreement, but rarely other researchers get different annotators to rate the same pairs. A notable exception is the Miller and Charles 30 (MC30) dataset [20], whose pairs were included in other datasets, namely RG65 [31], WS353 [8] and PS65 [29]. This provided us with the opportunity to compare how LLMs perform against different groups of annotators, all rating the same subset of pairs. For the pairs from MC30, we first gathered their scores from the four datasets. Then we asked several LLMs to rate the same pairs, repeating the queries 10 times. Finally, we measured the correlation between each LLM result and each dataset (using Pearson's correlation coefficient). The resulting confusion matrix is displayed in Table 2.

The average correlation between LLMs (the blue area of the table) is 0.92. The correlation between the LLMs and the datasets (pink area) is 0.90, and the average correlation between human-annotated datasets (yellow area) is 0.94. The fact that the LLM results correlate so highly with human annotators is promising and supports the hypothesis that LLMs can be used as a replacement for human annotators.

2.3 Prior knowledge of datasets

To assess the LLMs capabilities regarding semantic measures (Sections 2.1 and 2.2) we used datasets described in the literature. These datasets are, almost with no exception, published online and publicly available, either in the original paper, the authors website or, frequently, replicated in third party web pages or repositories. LLMs have been known to occasionally replicate data included in their training datasets, even leading to yet unresolved lawsuits, such as OpenAI's Codex reproducing publicly shared code [25] or ChatGPT reproducing literary works [10]. This poses an additional challenge when interpreting the scores returned by an LLM: do these scores actually encode some *meaning* or is the model just parroting pairs and values which were included in its training corpora?

In order to determine whether the models had previous knowledge of SM datasets, we devised 3 tests:

1. Given the name and date of publication of the dataset, ask the model to provide a sample of the pairs included in the dataset. Measure the percentage of correct pairs.
2. Given a sample of the pairs included in the dataset, ask the model to provide a different sample of pairs. Measure the percentage of correct pairs.
3. Given a sample of pairs in the dataset, ask the model to rate their semantic relation strength. Measure the percentage of values *matching exactly* the values in the dataset.

We repeated each test, for each model and each dataset, 10 times. Then we averaged the results and built a heat matrix for each test, which can be seen in Figure 1.

`DS Sample From DS Name` shows that generally, the models are aware of a few datasets, as they were capable of correctly producing pairs belonging to them when asked by their name. MC30 [20], PS65 [29], RG65 [31] and WS353 [8] are the least surprising: not only they are among the most well known and cited datasets, but they also overlap with each other, so naturally their results are correlated. The high values for BG1000k [3, 24], MEN3000 [4], SCWS2003 [15] and SL7576 [33] are harder to explain. Their size is above average, which could explain the good results; however, other large datasets (e.g. WORD19k [7], TR9856 [17]) did not achieve the same kind of results.

The `DS Sample From DS Sample` results are more uniform: the models were generally incapable of producing a sample of pairs from a dataset when given another sample. The most noticeable trend here are the above average results for the model `claude-3-opus`. These results could be explained by overlaps in datasets (the same pairs appearing in more than one dataset). However, the only meaningful occurrence that we know of happens with the previously mentioned pairs from MC30, and the datasets which contain them present results which are not significantly different from the other datasets.

`DS Values Exact Matches` provided the lowest results. Generally the models did not return values matching exactly the ones published in the datasets. The Atlasify240 [13], MiniMayo [28], SimLex999 [14] and SL7576 [33] datasets obtained the best scores.

LLMs seem to have very low knowledge of existing semantic measures datasets. Other than the few datasets which obtained the best results in `DS Sample From DS Name`, the models were incapable of producing pairs and values matching the ones present in datasets.

3 Discussion and ongoing work

We have demonstrated that prompt variations have impact on LLMs results (Section 2.1). The results obtained and described in Section 2 seem to indicate that LLMs can be used to predict semantic measure values. LLMs results for the MC30 subset correlate very highly with human annotators, as described in Section 2.2. This, however, should be taken with a grain of salt as the datasets containing those pairs are included in the few datasets whose name and content LLMs seem to be aware of, as shown in Section 2.3. We are currently extending this work as follows:

1. We are evaluating the semantic measures predictions for multiple LLMs and multiple datasets. The result will be a general measure of the correlation between LLMs predictions and human annotations, which will allow us to determine if (and which) LLMs are suitable to be used as a proxy for human judgement of semantic measures.
2. We are building a new gold standard dataset, focused on Portuguese affective words annotated by university students. This will be an entirely new dataset, making it impossible for it to have been included in the LLMs training corpora. Testing the LLMs with this new dataset will produce results which are guaranteed to be independent from prior knowledge.

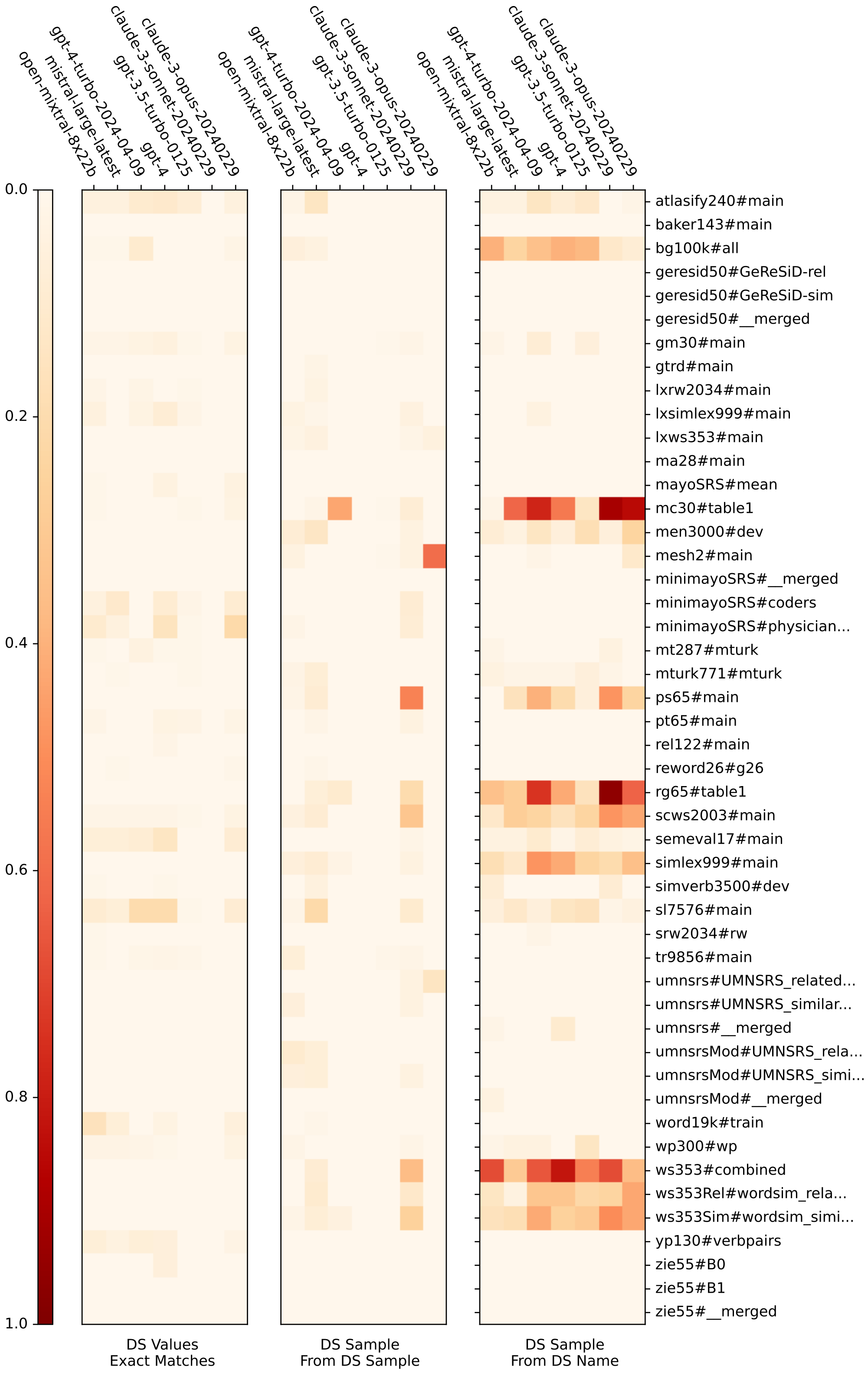

Figure 1 LLMs prior knowledge of semantic measures datasets.

A repository containing the datasets and the open source code to access them is available online[2]. The open source code we developed to query the models is still under development and is also available online[3].

If LLMs prove to be, in fact, a good measure of the *wisdom of the crowd* regarding semantic measures, they can then be used in several tasks which traditionally required human intervention or, at least, the use of human-generated datasets. For example, they can be used to evaluate semantic measure algorithms, or even to replace them.

References

1 Anthropic. Anthropic Models overview. `https://docs.anthropic.com/claude/docs/models-overview`. Accessed: May 8, 2024.

2 Anthropic. Tool use (function calling). `https://docs.anthropic.com/claude/docs/tool-use`. Accessed: May 9, 2024.

3 Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*, pages 29–36. Citeseer, 2006.

4 Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014. `doi:10.1613/JAIR.4135`.

5 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

6 Cohere. Models. `https://docs.cohere.com/docs/models`. Accessed: May 8, 2024.

7 Liat Ein Dor, Alon Halfon, Yoav Kantor, Ran Levy, Yosi Mass, Ruty Rinott, Eyal Shnarch, and Noam Slonim. Semantic relatedness of wikipedia concepts–benchmark data and a working solution. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

8 Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414, 2001. `doi:10.1145/371920.372094`.

9 Google AI for Developers. Gemini models. `https://ai.google.dev/gemini-api/docs/models/gemini`. Accessed: May 8, 2024.

10 Haleluya Hadero and David Bauder. New york times sues microsoft, open ai over use of content. *Globe & Mail (Toronto, Canada)*, pages B1–B1, 2023.

11 Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2023.

12 Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254, 2015. `doi:10.2200/S00639ED1V01Y201504HLT027`.

13 Brent Hecht, Samuel H Carton, Mahmood Quaderi, Johannes Schöning, Martin Raubal, Darren Gergle, and Doug Downey. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 415–424, 2012. `doi:10.1145/2348283.2348341`.

[2] `https://github.com/andrefs/punuy-datasets`

[3] `https://github.com/andrefs/punuy-eval`

14 Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015. `doi:10.1162/COLI_A_00237`.

15 Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 873–882, 2012. URL: `https://aclanthology.org/P12-1092/`.

16 Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048, 2023.

17 Ran Levy, Liat Ein Dor, Shay Hummel, Ruty Rinott, and Noam Slonim. Tr9856: A multi-word term relatedness benchmark. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 419–424, 2015. `doi:10.3115/V1/P15-2069`.

18 Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

19 Recalde Varela Pablo Marcel, Bolagay Egas Mauro Fernando, and Yanez Velasquez Jorge Roberto. A brief history of the artificial intelligence: chatgpt: The evolution of gpt. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5. IEEE, 2023.

20 George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

21 Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

22 Mistral AI. Mistral AI Large Language Models. `https://docs.mistral.ai/getting-started/models/`. Accessed: May 8, 2024.

23 Mistral AI Large Language Models. Function calling. `https://docs.mistral.ai/capabilities/function_calling/`. Accessed: May 9, 2024.

24 Sonya Nikolova, Jordan Boyd-Graber, and Christiane Fellbaum. Collecting semantic similarity ratings to connect concepts in assistive communication tools. *Modeling, Learning, and Processing of Text Technological Data Structures*, pages 81–93, 2012. `doi:10.1007/978-3-642-22613-7_5`.

25 Linda Novobilská. Free and open source software licensing requirements and copyright infringement involving artificial intelligence technologies. Master's thesis, Humboldt-Universität zu Berlin, 2023.

26 OpenAI API. Function calling. `https://platform.openai.com/docs/guides/function-calling`. Accessed: May 9, 2024.

27 OpenAI API. OpenAI Platform Documentation. `https://platform.openai.com/docs/models/overview`. Accessed: May 8, 2024.

28 Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299, 2007. `doi:10.1016/J.JBI.2006.06.004`.

29 Giuseppe Pirró and Nuno Seco. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In *On the Move to Meaningful Internet Systems: OTM 2008: OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part II*, pages 1271–1288. Springer, 2008. `doi:10.1007/978-3-540-88873-4_25`.

30 Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346, 2011. `doi:10.1145/1963405.1963455`.

31 Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965. doi:10.1145/365628.365657.

32 Hansen A Schwartz and Fernando Gomez. Evaluating semantic metrics on tasks of concept similarity. In *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*, pages 324–340. IGI Global, 2012.

33 Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, 2014. doi:10.3115/V1/P14-1068.

34 Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023. doi:10.1109/JAS.2023.123618.