Evaluating Large Language Models' Understanding of Financial Terminology via Definition Modeling

James Jhirad[1] **Edison Marrese-Taylor**[2,3] **Yutaka Matsuo**[3]
[1] Faculty of Arts & Science, University of Toronto
[2] National Institute of Advanced Industrial Science and Technology
[3] Graduate School of Engineering, The University of Tokyo
james.jhirad@mail.utoronto.ca,{emarrese,matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

As machine learning models grow increasingly sophisticated, the question of their reliability and understanding becomes critical, especially within specialized domains like finance, medicine, and law. Definition modeling, the task of generating a textual definition from a word, has proven a useful technique to help better understand word sense and embeddings, which is at the core of language learning and word acquisition. Drawing upon a repository of financial terminology (Investopedia), we build a dataset of 14,000 terms and definitions. We design a number of tasks to evaluate the capacity of various LLMs to generate accurate and expansive definitions in this domain, and propose to utilize definition modeling to probe the abilities of Large Language Models (LLMs) in the context of the financial domain. Using our dataset, we present an empirical study where we test a broad selection of LLMs on our tasks with zero-shot and k-shot approaches. We show the extent to which these models are able to define financial terms. We observe large performance increases for smaller models with in-context learning, and see that they are almost comparable to GPT-3.5. Our work shows the boons of using definition modeling to evaluate models in more specific fields of study, such as finance.

1 Introduction

Definition modeling is the task of estimating the probability of a textual definition, given a word being defined. Since its conception (Noraset et al., 2017) several approaches have tackled this task, and over the past few years have achieved substantial performance improvements. This task has been shown to give an arguably more transparent view of the extent to which syntax and semantics are captured by a model.

So far, existing approaches for this task have followed the traditional approach, where models are trained on a corpus of word-definition pairs to later be tested on how well they generate definitions for words not seen during training. However, the recent success of Large Language Models (LLMs) has caused a shift in our field, showing that such models can achieve excellent performance on a wide variety of downstream tasks, utilizing zero-shot or few-shot approaches (Brown et al., 2020; Kojima et al., 2022), i.e. without fine-tuning.

Term	Definition (1st Key Takeaway)
Enterprise value (EV)	Enterprise value (EV) measures a company's total value, often used as a more comprehensive alternative to equity market capitalization.
Bonds	Bonds are units of corporate debt issued by companies and securitized as tradeable assets.

Table 1: Examples of Term-Definition pairs taken from dataset built out of Investopedia

Furthermore, while the definition modeling task was originally intended for the study of *general* words, we note that some recent work has focused on extending the task to domain specific terms. These works propose domain specific models trained to define specialized terminology. While these efforts are welcome, and remain an interesting and useful approach, we note that they are still limited in terms of scope, with key domains such as finance or chemistry, being left out so far.

In this paper, we tackle the two aforementioned issues by: (1) Using definition modeling tasks as a probe to test the abilities of LLMs in a zero-shot or few-shot setting, motivated by the original ideas of (Noraset et al., 2017), and (2) Introducing a new dataset with 14,000 term-definition pairs in the financial domain - so far an unexplored direction which we believe is particularly relevant due to the way in which LLMs are trained, while also

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 93–100
November 1–4, 2023.

presenting relevant use-cases.

To this end, we construct a new dataset of approximately 14,000 terms specific to the financial domain. Table 1 shows examples of how our data looks like. We use this new dataset to assess the quality of financial definitions generated from some of the latest LLMs with a zero-shot and few-shot approach. Our proposals offer a concrete direction for prompting methodology, examining variables such as number of shots, usage of word context, role of domain, and evaluation.

Our experiments show that while some of the larger and most cutting-edge LLMs are able to define financial terms, they outperform classical definition modeling tasks. Our choice of COMET as an evaluation metric also appears like an adequate metric to use for definition modeling tasks. Lastly, we release our code to encourage research in this direction[1].

2 Related Work

Our work is primarily related to the seminal work by Noraset et al. (2017) and Hill et al. (2016), in which a model is tasked with generating a definition for a word given its respective embedding, or with mapping dictionary definitions to lexical representations of words, respectively.

After this, several works have proposed improvements. Many introducing techniques and datasets to address several shortcomings of the initial ideas. For example, Gadetsky et al. (2018) addresses polysemy and presents a dataset from Oxford Dictionaries, where each definition is also supplemented with context sentences, in which each example word is used, allowing models to disambiguate. Ni and Wang (2017) proposed an approach for automatically explaining slang English terms in a sentence, and introduced yet another dataset. Ishiwatari et al. (2019) and Reid et al. (2020) propose to further rely on local and global contexts, with the latter also introducing a dataset based on Cambridge Dictionaries, and a dataset for French.

More recently, Huang et al. (2021) study the problem of definition specificity, and propose a method for tuning a model to account for hyper focused (over-specific) or highly general (under-specific) definitions. Chen and Zhao (2022) propose to unify the seminal ideas of reverse dictionary and definition modeling in a single model, with the goal of helping better understand word sense and embeddings.

Finally, we find several works aiming to generate definitions in specialized fields, whose efforts are well aligned with our work. August et al. (2022) propose to generate definitions of scientific and medical terms with varying complexity, using a dataset constructed from consumer medical questions and science glossaries (MedQuAD and Wikipedia). Liu et al. (2021) introduce Graphine, a dataset for biomedical terminology definition, and Huang et al. (2022) propose to model 'jargon', with a dataset constructed semi-automatically based on Wikipedia and Springer. Though our work is similar, since we also extend definition modeling to a new domain, critically our approach differs as we 'probe' our models' understanding of financial terms through definition modeling.

3 The INVESTOPEDIA Dataset

To construct a dataset of financial terms, we rely on Investopedia, an extensive repository of financial information and terminology. We initiated our data collection by fetching every available link on the website. Out of approximately 25,000 extracted links, half were found to contain articles or news discussing specific events, while the remaining comprised pages focused on delineating specific financial terms or concepts.

The extraction of the main term from the article presented a complex task due to the variety in article title structures. To achieve this, we employed a set of heuristics that guided the selection of the most appropriate string as the main term, which were also combined with manual annotation.

A similar strategy was used for identifying potential acronyms associated with each term, incorporating both heuristic and manual processes.

Investopedia Dataset	
# of Terms	13,609
Definition Length (in tokens)	19.69 ± 7.33
Mean # Key Takeaways	3.66
Oxford Dataset	
# of Terms	122,319
Definition Length (in tokens)	11.03 ± 6.97

Table 2: Statistics on our financial data and the Oxford dataset. The definitions for the financial data is considered to be the first key takeaway.

[1]https://github.com/KobiJames/Financial-DefMod

The second phase of our data scraping involved the extraction of the "Key Takeaways" section found in each article. This section, typically consisting of 3 to 5 bullet points summarizing the article, is a critical part of our dataset. We chose to leverage this section over the first paragraph to generate the definitions for each term. The choice of the "Key Takeaways" section over the first paragraph was primarily due to its structure and ease of extraction. While both sections offered a general description of the term, the former provided something more akin to a set of distinct 'mini definitions'. Each point in the "Key Takeaways" is grammatically independent of the previous, and captures something different about the article than the other points. This encapsulation of the entire article's content in conveniently parsed pieces, made it an ideal source for our definitions.

Our finalized dataset consists of data points, each corresponding to a single article. The primary components of each data point include the processed term name, any associated acronym, and a list of key takeaways serving as the definition. Supplementary data, added for convenience and potential further research, encompasses the article's header, and the full text of the article, separated by HTML tags ('h1','h2','p', etc...) and ordered from top to bottom in an array.

4 Empirical Study

4.1 Experimental Setup

To test the abilities of LLMs in performing definition modeling in the financial domain, we leverage the "Key Takeaways" section in INVESTOPEDIA. Based on the success of previous work in augmenting definition modeling with contextual information (Gadetsky et al., 2018; Ishiwatari et al., 2019; Reid et al., 2020), we additionally leverage sample phrases from the article content, which we present to the models as context in addition to the term, for generating the definition. They are presented to our models by appending them to the prompt, and we refer to them as "examples" in further discussion.

The structure and wording of the prompts presented to the models were established through a systematic testing process. Most notably, we observed that prompting the models with the task to define a "financial term" led to consistently significant performance improvements. A similar improvement was seen when the models were asked to present their responses as points or bullet points, leading us to incorporate these findings into our final prompt, which we used across all the experimental runs in the zero-shot scenario. For our experiments incorporating examples or for the k-shots settings, we append each phrase or shot above the prompt, and add a newline character to separate the base prompt, examples, and shots. For prompts with both examples and shots, shots are put above examples. For prompts with examples, we use two examples, and for prompts with shots, we use two shots. Please see our supplementary material for details.

Regarding model choice, we consider a broad section of LLMs varying widely in terms of the number of parameters and training scheme. Concretely, we work with the following models: OPT-IML 1.3B (Iyer et al., 2023), GPT-J (6B parameters) (Wang and Komatsuzaki, 2021), GPT-JT (6B parameters) (Rel, 2022), MT0-XXL (13B parameters), FLAN-UL2 (20B parameters) (Tay et al., 2023; Muennighoff et al., 2023), and ChatGPT/GPT-3.5 Turbo (175B parameters).

In terms of evaluation, we note that previous work has mainly utilized n-gram overall metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). Reid et al. (2020) showed that by migrating to metrics based on Machine Learning (ML), such as BERTScore (Zhang et al., 2019), it was possible to more adequately capture nuance in the definitions generated by their approach, which led more recent models to adopt this evaluation metric as well. In this paper, we take another step in this direction and experiment with COMET (Rei et al., 2020) as an evaluation metric for our task. Though COMET is intended for evaluation of Machine Translation models, in our early experiments we found that it offered a robust alternative to other ML-based metrics such as BERT-Score and worked well at capturing the ability of the models at generating accurate and expansive definitions.

We also note that the complexity of our terms, evidenced by the length and detail of the "Key Takeaways" section of our dataset, pose a considerable challenge in our evaluation process. We observe that any of the takeaway points can, to a degree, serve as a separate definition of the term. To avoid penalizing the model when it generates such relevant information, we propose a scoring scheme where we compare the output of the model against combinations of "Key Takeaways". Let x_i be our target financial term. For $i \in [1 \dots L]$, we have $j \in [1 \dots N_i]$, where L is the size of our dataset,

Data	Model	Params.	COMET				METEOR				BLEU			
			base	ex.	sh.	sh.+ex.	base	ex.	sh.	sh.+ex.	base	ex.	sh.	sh.+ex.
Investopedia (1st KTAW)	OPT-IML	1.3B	47.82	47.41	59.75	60.01	15.92	17.11	25.96	29.74	4.54	5.26	6.48	6.36
	GPT-J	6.0B	50.77	43.73	57.55	60.80	22.17	16.94	28.32	32.37	3.56	3.30	4.97	6.52
	GPT-JT	6.0B	33.23	44.41	57.18	60.35	6.65	18.60	27.99	31.16	0.96	3.33	4.44	5.85
	MT0-XXL	13B	55.21	64.55	63.06	66.10	24.68	30.18	24.82	28.85	4.90	7.81	6.72	8.95
	FLAN-UL2	20B	66.30	68.49	68.15	<u>68.98</u>	27.30	30.17	30.16	32.31	7.93	10.16	9.73	10.66
	GPT-3.5-Turbo	175B	68.64	69.12	68.51	**<u>69.79</u>**	35.82	37.29	36.63	38.22	5.74	6.36	6.73	7.15
Investopedia	OPT-IML	1.3B	48.45	48.92	61.21	62.95	16.38	18.48	27.33	32.67	2.99	4.28	5.65	7.06
	GPT-J	6.0B	52.11	45.27	59.93	63.77	23.56	18.39	29.86	34.97	3.93	3.81	5.51	7.55
	GPT-JT	6.0B	33.99	45.98	59.61	63.68	7.12	20.49	29.62	34.39	1.01	3.76	4.98	7.01
	MT0-XXL	13B	56.43	67.47	63.69	68.10	26.12	33.28	25.40	31.00	4.38	7.17	3.85	6.46
	FLAN-UL2	20B	66.63	69.74	68.67	<u>70.61</u>	27.71	31.72	30.68	34.19	4.11	6.50	5.82	7.44
	GPT-3.5-Turbo	175B	72.18	72.82	71.98	**<u>73.57</u>**	38.50	40.02	39.38	41.20	6.65	7.38	7.78	8.32
Oxford	OPT-IML	1.3B	32.66	32.88	47.23	49.02	2.78	2.65	7.80	9.09	0.29	0.39	3.29	3.43
	GPT-J	6.0B	39.00	40.61	38.20	42.80	8.14	10.49	10.45	13.21	0.84	1.12	1.26	1.28
	GPT-JT	6.0B	40.88	40.59	39.54	42.64	9.22	10.18	12.10	14.40	0.97	1.24	1.38	1.61
	MT0-XXL	13B	46.08	50.39	33.75	39.45	8.70	9.75	3.32	8.50	2.80	3.35	0.50	1.40
	FLAN-UL2	20B	45.72	50.10	47.52	<u>50.46</u>	6.73	9.83	8.41	10.70	2.63	4.03	3.40	4.40
	GPT-3.5-Turbo	175B	55.93	59.69	57.96	**<u>61.75</u>**	23.90	27.99	25.26	29.93	3.57	4.74	8.63	9.84

Table 3: Scores across COMET, METEOR and BLEU on financial definition modeling tasks and the Oxford dataset on "dictionary" definition modeling, where we use. 'ex.' to indicate use of examples, 'sh.' to indicate use of 2-shot, and KTAW is short for "Key Takeaway". Values which are bold & underlined are the best out of all the models; values that are only underlined are the best results from open-source models

and N_i denotes the number of "Key Takeaways" for term x_i. Let $y_{i,j}$ be the jth "Key Takeaway" for x_i, and the set $\{y_{i,j}|j \in [1 \dots N_i]\}$ is the list of sorted "Key Takeaways" to be used as targets. Finally, let $P(s)$ be the function that generates the power set of a set s. Then, for a given evaluation Metric M, our final score is computed following Equation 1, below, where $a; B$ denotes the sequential string concatenation of every element in B to the end of a in order (i.e. $a + b_1 + b_2 + \dots$).

$$S_i := \max\{M(\hat{y}, [y_{i,1}; K])|K \in P([y_{i,2}, \dots y_{i,N_i}])\} \quad (1)$$

The idea of combining "Key Takeaways" in such a way derives from the fact that the first bullet point typically provides the most straightforward definition of the term. By combining this with the rest of the points, we ensured that a wide array of definitions, ranging from simplistic to comprehensive, were accounted for in our evaluation scheme.

While this method of evaluation is holistic, it also makes it difficult to compare results of our evaluations against different datasets that lack this type of information for their gold standard. Evaluation scores could possibly be inflated, which would cause an uneven comparison. Evaluations over our dataset just using the first key takeaway as a gold standard were added as a baseline for the purpose of comparability. Finally, to understand the complexity of defining financial terms, we also test our approach on the dataset constructed from Oxford dictionaries (Gadetsky et al., 2018). To the best of our knowledge, our work is the first one to test the abilities of LLMs on the "dictionary" definition modeling task using zero-shot or few-shot settings. We believe results on this will help contextualize our results in the financial domain, while also giving new insight into the abilities of LLMs.

4.2 Prompts

The following is the prompt structure we used for each term, as well as every variation of the prompt we used for the term "*A-B Trust*".

$$T \in \textbf{Investopedia Terms}$$

$$E_i^T \in \textbf{Examples for term } T|i \in 0, 1$$

$$S_i \in \textbf{Shot List}|i \in 0, 1$$

Prompt Structure

"S_0
S_1
E_0^T
E_1^T
Define, in a financial context, '$[T]$' with bullet points. Please list up to 2 bullet points.

Definition: ”
Prompt without examples or shots
“Define, in a financial context, ‘A-B trust’ with bullet points. Please list up to 2 bullet points.
Definition: ”
Prompt with examples
“06 million will opt for an A-B trust in 2022
While A-B trusts are a great way to minimize estate taxes, they are not used much today
Define, in a financial context, ‘A-B trust’ with bullet points. Please list up to 2 bullet points.
Definition: ”
Prompt with shots
“Define, in a financial context, ‘Enterprise Value’ with bullet points. Please list up to 2 bullet points.
Definition: Enterprise value (EV) measures a company’s total value, often used as a more comprehensive alternative to equity market capitalization. Enterprise value includes in its calculation the market capitalization of a company but also short-term and long-term debt and any cash on the company’s balance sheet.
Define, in a financial context, ‘Bond’ with bullet points. Please list up to 2 bullet points.
Definition: Bonds are units of corporate debt issued by companies and securitized as tradeable assets. A bond is referred to as a fixed-income instrument since bonds traditionally paid a fixed interest rate (coupon) to debtholders.
Define, in a financial context, ‘A-B trust’ with bullet points. Please list up to 2 bullet points.
Definition: ”
Prompt with examples and shots
“Define, in a financial context, ‘Enterprise Value’ with bullet points. Please list up to 2 bullet points.
Definition: Enterprise value (EV) measures a company’s total value, often used as a more comprehensive alternative to equity market capitalization. Enterprise value includes in its calculation the market capitalization of a company but also short-term and long-term debt and any cash on the company’s balance sheet.
Define, in a financial context, ‘Bond’ with bullet points. Please list up to 2 bullet points. Definition: Bonds are units of corporate debt issued by companies and securitized as tradeable assets. A bond is referred to as a fixed-income instrument since bonds traditionally paid a fixed interest rate (coupon) to debtholders.
06 million will opt for an A-B trust in 2022
While A-B trusts are a great way to minimize estate taxes, they are not used much today
Define, in a financial context, ‘A-B trust’ with bullet points. Please list up to 2 bullet points.
Definition: ”

4.3 Results

The performance outcomes for all selected models, based on INVESTOPEDIA are displayed in Table 3. We also show our results on the classic or “dictionary” definition modeling task, for which all models were evaluated under comparable configurations. To assess the performance of the models, three evaluation metrics were employed, namely COMET, METEOR, and BLEU, with focus on the former. We use four different prompting methods, as described in section 4.1; ‘base’ for no shot, no examples, ‘ex.’ for no shot, 2 examples, ‘shot’ for 2 shot, no examples, and ‘shot+ex.’ for 2 shots, 2 examples.

As shown in Table 3 we observe that the performance of all considered models aligns with respective size — the larger the model, the better the performance. We also note that smaller models seem to be unable to perform well without the help of the additional context presented. Finally, the performance gain of smaller models with in-context learning is substantial relative to the baselines.

In our exploration of the effects of differing prompting styles, an interesting pattern emerged, particularly when contrasting the smaller models with larger ones. The performance of MT0-XXL, Flan-UL2, and GPT-3.5-Turbo increase when augmented with examples in the prompt (i.e. from base to ex., or sh. to sh.+ex.), while OPT-IML, GPT-J, and GPT-JT are unaffected by this change. This discrepancy is also seen in the results on the Oxford dataset, further exemplifying the relevance of this observation. Since, MT0-XXL, Flan-UL2, and GPT-3.5-Turbo are larger models. This raises a compelling hypothesis that, for models without direct knowledge of terms, model size may be correlated with enhanced capabilities of utilizing non-shot context to increase performance. This is seen by the aforementioned model’s increase in performance with examples over our other models.

Results on the Oxford dataset exhibit a decrease in performance relative to the financial definition modeling tasks. We ascribe this primarily to the discrepancy in length between the Oxford gold standard and the model outputs, as can be seen in Table 2. The brevity of the Oxford definitions contrasts with the verbose model responses, contributing to the performance gap for GPT-3.5-turbo on a seem-

Model	COMET	METEOR	BLEU
INVESTOPEDIA			
Ours (FLAN-UL2)	68.98	32.31	10.66
Oxford			
Gadetsky et al. (2018)	-	-	23.77
Ishiwatari et al. (2019)	-	-	25.19
Reid et al. (2020)	57.00	35.05	27.38
Huang et al. (2021)	-	-	26.52
Ours (GPT-3.5-Turbo)	61.75	29.93	9.84

Table 4: Comparison of our best performing models against state-of-the-art approaches for the Oxford dataset.

ingly simpler task. While we intend for Oxford to be a type of baseline for our tests, INVESTOPEDIA has more expansive definitions, which make comparisons using COMET difficult. For comparison, our dataset has around three to four key points for each definition, each, at minimum, the length of a definition from the Oxford dataset. This discrepancy exposes a limitation of the COMET evaluation metric, despite its many advantages.

Concerning results of our special evaluation schema which uses multiple gold standards, compared to only using the first key takeaway, we see an across the board improvement of each model by a few comet points. This is expected, as taking a max across multiple gold standards will inevitably boost evaluation scores.

Finally, we compare our best results on INVESTOPEDIA and the Oxford benchmark against state-of-the-art models for the latter, all based on fine-tuning. Specifically, we consider the approaches by Gadetsky et al. (2018) who released the Oxford dataset, Ishiwatari et al. (2019) who proposed a local-and-global context model based on word embeddings, Reid et al. (2020) who leveraged BERT (Devlin et al., 2019) and combined it with a variational inference framework, and Huang et al. (2021) who propose a specificity-sensitive approach with models based on T5 (Raffel et al., 2020).

As Table 4 shows, we see that finetuning-based approaches are able to outperform our k-shot and zero-shot techniques based on prompting by a large margin in terms of BLEU. We also see that the overall best performance of the latter techniques on INVESTOPEDIA remain on pair with Oxford in terms of all metrics, suggesting that fine-tuning could also lead to improved results in our dataset as well. We think this could be interesting for specific applications where generating accurate definitions of financial terms is needed.

5 Conclusions

In summary, this paper advances our understanding of the capabilities and limitations of Large Language Models (LLMs) in specialized domains by using definition modeling tasks as a lens into our models abilities. We have established a framework for testing LLMs in a zero-shot or few-shot setting and demonstrated the utility of this approach with a novel dataset of 14,000 term-definition pairs in the financial domain - an area so far underrepresented in such studies. Our empirical results, derived from an array of LLMs, highlight the degree to which these models can define financial terms accurately and expansively. Notably, we observed considerable performance enhancements when adding in-context learning to smaller models, indicating that they can approach the performance levels of larger counterparts, such as GPT3.5. While our study presents a significant step towards comprehending the nuances of LLMs in specialized areas like finance, it also underscores the challenges that remain. The performance gap witnessed on tasks with shorter definitions, like those in the Oxford dataset, reveals inherent limitations of the evaluated models and evaluation metrics. We hope our work inspires further research into the application of definition modeling as a means to understand and refine LLMs, particularly in critical fields such as finance, law, and medicine.

Limitations

There are a few notable limitations of our work. Firstly, the methods we used were primarily in-context learning. We did not fine-tune any models, although this was by intention. While our goal was to 'probe' the models we chose, it does remain a question whether our dataset can be used for performance gains with training. We leave this to future work.

Ethics Statement

Our main objective is to propose a new task to evaluate the abilities of LLMs, introducing a dataset of financial term definition. One potential use-case is to have a model generate fake definitions that may mislead users that interact with an LLM when deployed. By publicly releasing our data, we hope to minimize such risks.

Acknowledgements

For our experiments, computational resources of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) were used. We are also grateful to the NVIDIA Corporation, which donated one of the GPUs used for this research.

References

2022. Releasing GPT-JT powered by open-source AI.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating Scientific Definitions with Controllable Complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pinzhen Chen and Zheng Zhao. 2022. A Unified Model for Reverse Dictionary and Definition Modelling. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 8–13, Online only. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition Modelling for Appropriate Specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. Understanding Jargon: Combining Extraction and Generation for Definition Modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4004, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to Describe Unknown Phrases with Local and Global Contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*.

Zequn Liu, Shukai Wang, Yiyang Gu, Ruiyi Zhang, Ming Zhang, and Sheng Wang. 2021. Graphine: A Dataset for Graph-aware Terminology Definition Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3453–3463, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao,

M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning.

Ke Ni and William Yang Wang. 2017. Learning to Explain Non-Standard English Words and Phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *AAAI Conference on Artificial Intelligence*. AAAI Conference on Artificial Intelligence.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. VCDM: Leveraging Variational Bi-encoding and Deep Contextualized Word Representations for Improved Definition Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6331–6344, Online. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.