NutFrame: Frame-based Conceptual Structure Induction with LLMs

Shaoru Guo[1,2], **Yubo Chen**[2,3*], **Kang Liu**[2,3,4], **Ru Li**[1], **Jun Zhao**[2,3]
[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China
[2]The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China
{shaoru.guo, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn, liru@sxu.edu.cn

Abstract

Conceptual structure is fundamental to human cognition and natural language understanding. It is significant to explore whether Large Language Models (LLMs) understand such knowledge. Since FrameNet serves as a well-defined conceptual structure knowledge resource, with meaningful frames, fine-grained frame elements, and rich frame relations, we construct a benchmark for coNceptual structure induction based on FrameNet, called **NutFrame**. It contains three sub-tasks: Frame Induction, Frame Element Induction, and Frame Relation Induction. In addition, we utilize prompts to induce conceptual structure of Framenet with LLMs. Furthermore, we conduct extensive experiments on NutFrame to evaluate various widely-used LLMs. Experimental results demonstrate that FrameNet induction remains a challenge for LLMs.

Keywords: FrameNet, Frame Induction, Frame Element Induction, Frame Relation Induction

1. Introduction

Large Language Models (LLMs) have exhibited impressive performance on most natural language processing tasks (OpenAI, 2023, 2022; Chowdhery et al., 2023; Thoppilan et al., 2022). This has led to a recent surge in studies to explore the extent of knowledge within LLMs. Existing studies mainly focus on syntactic knowledge (Liu et al., 2019; Hu et al., 2020) and world knowledge (Liu et al., 2021; Peng et al., 2022; Petroni et al., 2019). However, the extent to which these models reflect the human-like cognitive abilities to extract structured representations of concepts is not well-understood (Patterson et al., 2007; Collins and Olson, 2014).

Conceptual structure refers to the way concepts are organized, represented, and interconnected in the human mind (Smoliar, 1987; Guo et al., 2023). When human beings experience the world, they conceptualize their experiences into concepts, and organize them into a highly complex and hierarchical structure through the brain rather than being stored randomly (de Beaugrande and Dressler, 1986). For example, when the word "buy" is given, people recall information from their memory and activate the concept `Commerce_buy`, which includes properties like "buyer", "goods", "money", and more. Moreover, the concept `Commerce_buy` is organized into a structure with relations, such as `Commerce_goods-transfer` $\xrightarrow{Perspective\_on}$ `Commerce_buy`, indicating that `Commerce_buy` is a fundamental scene of `Commerce_goods-transfer` from the perspective of the buyer.

FrameNet (Baker et al., 1998; Fillmore, 1976)

*Corresponding author

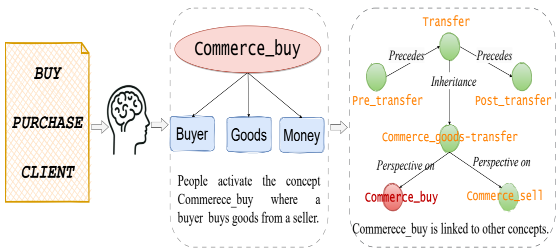

Figure 1: An Example of Conceptual Structure.

is an excellent repository of conceptual structure knowledge designed by experts. Typically, each sense of a word belongs to a frame, which is a conceptual structure that describes a particular type of entity or event and the participants involved therein (called frame elements, FEs). Moreover, FrameNet also provides Frame-to-Frame relations (Guan et al., 2023). As shown in Figure 1, frame `Commerce_buy` about "buy" involves the FEs "buyer", "goods" and "money". Moreover, FrameNet organizes frames into a net through rich Frame-to-Frame relations, such as `Pre_transfer` $\xrightarrow{Precedes}$ `Transfer`.

Thus, we comprehensively evaluate the ability of LLMs to induce FrameNet conceptual structure by designing three tasks: (1) **Frame Induction (FI)** task aims to induce the meaningful frames. Given a set of lexical units or a description, the FI task requires LLMs to induce the corresponding frame. For example, given lexical units such as "buy", "client" and "purchase", the FI task aims to induce the frame `Commerce_buy`. (2) **Frame Element Induction (FEI)** task aims to induce fine-grained frame elements associated with frames. Given

LREC-COLING 2024, pages 12330–12335

the frame `Commerce_buy`, the FEI task requires LLMs to induce its frame elements, such as "buyer", "money", "goods" and so on. (3) **Frame Relation Induction (FRI)** task aims to organize frames with rich frame relations. Given the frames `Transfer` and `Commerce_goods-transfer`, the FRI task aims to predict the "Inheritance" between them.

Based on the aforementioned considerations, we construct a benchmark for coNceptual structure induction based on FrameNet called *NutFrame*. We use prompts to induce conceptual knowledge with LLMs. Furthermore, we conduct extensive experiments on NutFrame to evaluate the ability of widely-used LLMs, including GPT-4 (OpenAI, 2023), ChatGPT (OpenAI, 2022), Llama-2 (Touvron et al., 2023), and ChatGLM (Du et al., 2022; Zeng et al., 2023). The experimental results demonstrate that FrameNet induction is still a challenge for LLMs.

- We propose a systematic study to induce Frame-based conceptual structure knowledge with LLMs, which is highly valuable yet has been ignored by previous works.
- We construct *NutFrame*, a benchmark for coNceptual structure induction based on FrameNet. Additionally, we use prompts to induce FrameNet with LLMs and devise evaluation metrics to assess the ability.
- We conduct extensive experiments on NutFrame with widely-used LLMs, including GPT-4, ChatGPT, Llama-2, and ChatGLM. The experimental results show that FrameNet induction remains a challenge for existing LLMs.

2. NutFrame

In this session, we introduce the dataset construction process of our NutFrame, which consists of three sub-datasets: Frame Induction, Frame Element Induction, and Frame Relation Induction.

2.1. Frame Induction Dataset

Frame Induction (FI) aims to leverage LLMs to induce frames using lexical units or descriptions. We construct FI data from two aspects: Lexical_Unit-based FI dataset and Description-based FI dataset.

Lexical_Unit-based FI dataset. The frame represents shared semantics of lexical units in a way that is easily understandable to humans. Thus, we extract lexical units and their frames from FrameNet and then organize them into pairs. For example, a lexical_unit-frame pair such as "<*buy, client, purchase... || Commerce_buy*>" is created. The lexical_Unit-based FI contains 1,073 pairs, as shown in Table 1.

Task	Dataset	Number
FI	Lexical Units	13,640
	Lexical Unit-based FI[♡]	1,073
	Description-based FI[♡]	1,221
FEI	Frame Element	11,428
	FEI[♡]	1,221
FRI	Frame Relation	8
	FRI[♡]	1,849

Table 1: Statistics of the NutFrame dataset. ♡ represents the number of pairs constructed in this work.

Description-based FI dataset. Descriptions are more flexible for representing frames and are more informative[1]. Thus, we extract the frames and their descriptions from FrameNet and organize them into pairs. For example, "<*A buyer and a seller exchange money and goods... || Commerce_buy*>" is a description-frame pair. The description-based FI consists of 1,221 frame-description pairs, as shown in Table 1.

2.2. Frame Element Induction Dataset

Frame Element Induction (FEI) aims to leverage LLMs to induce frame elements for given frames.

Frame elements are semantically defined roles that are specific to a frame. Thus, we extract frames along with their elements from FrameNet and organize them into pairs. For instance, a frame-frame_element pair could be represented as "<*Commerce_buy || buyer, goods, money...*>". The FEI consists of 11,428 frame elements, with an average of 9.35 elements assigned to each frame, as shown in Table 1.

2.3. Frame Relation Induction Dataset

Frame Relation Induction (FRI) aims to leverage LLMs to predict relations for given frames.

We introduce FRI, a framework designed to predict relations between two frames. To achieve this, we extract frames and relations from FrameNet. These frames are then converted into sequences, which are combined with their corresponding relation types. For example, the "<*Pre_transfer, Transfer || Precedes*>" exemplifies such a frame sequence and relation type. The FRI consists of 1,849 pairs, as shown in Table 1.

[1]FrameNet includes non-lexical_unit frames that establish connections between frames in specific scenarios, such as `Commerce_goods-transfer`. In this particular situation, inducing frames solely based on lexical units becomes unfeasible; however, induction based on descriptions remains possible.

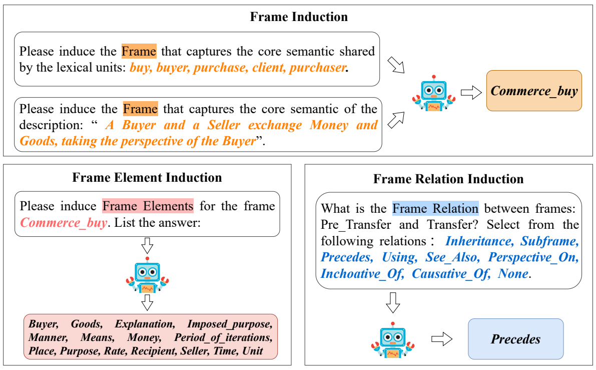

Figure 2: Illustration for FrameNet Induction.

3. FrameNet Induction Method

In this section, we introduce the methods used to explore FrameNet semantic knowledge in LLMs.

3.1. Frame Induction Method

FI focuses on evaluating the ability of LLMs to induce frames. As depicted in Figure 2, LLMs are expected to induce the frame `Commerce_buy` using the provided set of lexical units or descriptions. To achieve this, we employ prompts such as "*Please induce the frame that captures the core semantics shared by the [frame.lexical units] or [frame.description]*" for frame induction.

3.2. Frame Element Induction Method

FEI focuses on evaluating the ability of LLMs to induce frame elements. As illustrated in Figure 2, LLMs should be capable of inducing frame elements, such as "buyer" and "goods", for the frame `Commerce_buy`. Therefore, we utilize prompts such as "*Please induce the frame elements for the [frame.name]*" to query the LLMs for frame elements associated with a particular frame.

3.3. Frame Relation Induction Method

FRI focuses on evaluating the ability of LLMs to predict the relations between frames. As shown in Figure 2, given frames `Pre_transfer` and `Transfer`, LLMs are expected to predict the "Precedes", as it indicates that `Pre_transfer` occurs before `Transfer`. Thus, we utilize prompts such as "*Please identify the relation between [frame.name1, frame.name2] and select the relation type from the options: [frame.relation]*" to predict the relation type[2].

[2]We have introduced a NONE category to represent the absence of any relation between frames.

4. Experiments

In this section, we introduce experiment setup, and then report the results and analysis.

4.1. Experiment Setup

Models. We experiment with several LLMs, including GPT-4, ChatGPT, Text-Davinci-003, Llama-2 (7B), ChatGLM (6B), and GLM (130B).

Evaluation Metrics. For FI, we use Mean Reciprocal Rank (MRR) and Hits@k (Yang et al., 2012). For FEI, we use Micro-F1 and Macro-F1. For FRI, we employ precision, recall, and F1-score (Sakaguchi et al., 2021).

4.2. Main Results

From Table 2, 3 and 4, we can conclude that:

(1) **FrameNet induction presents a challenge for LLMs**. The poor performance of LLMs across three induction tasks, such as 32.61% Hits@5 for FI, 41.73% Micro-F1 for FEI, and 26.32% F1-score for FRI, indicates that existing LLMs have difficulties in inducing FrameNet. This may be attributed to the implicit nature of FrameNet within texts.

(2) **GPT-4 outperforms other LLMs**. Taking description-based FI as an example in Table 2, GPT-4 achieves 32.61% Hits@5, surpassing ChatGPT (23.53%) and other baselines.The same conclusion applies to FEI (Table 3) and FRI (Table 4).

(3) **Few-shot Learning outperforms Zero-shot Learning**. For example, in Table 3, few-shot learning achieves +24.11% improvement compared to zero-shot learning on FEI (41.73% vs. 17.62%).

4.3. Results of Frame Induction

LLMs tend to generate concrete frames, lacking the desired ability of abstraction. For example, when provided with the lexical units or description of `Commerce_scenario`, LLMs always generate the `Commerce_goods-transfer`[3].

LLMs provided with descriptions outperform those relying on lexical units. As shown in Table 2, description-based FI consistently outperforms lexical_unit-based FI. This may be because description offers more contextual informations.

4.4. Results of Frame Element Induction

Frame elements generated by LLMs are more general and not specific to each frame. For example, LLMs tend to generate general FEs like "agent" and "theme" when provided with the `Standing_by` frame. In contrast, human experts are able

[3]It is worth noting that `Commerce_goods-transfer` is considered more concrete as it is a subframe of the `Commerce_scenario` in FrameNet.

Method	Zero-shot(%)				Few-shot(%)			
	MRR	Hits@1	Hits@3	Hits@5	MRR	Hits@1	Hits@3	Hits@5
Lexial_Uints-based Frame Induction								
Llama-2 (7B)	1.48	0.83	2.54	3.22	8.76	7.12	10.27	11.27
ChatGLM (6B)	1.75	0.66	2.15	2.82	8.22	6.79	9.36	10.57
GLM (130B)	4.93	4.08	5.80	6.21	14.60	12.18	16.82	18.31
Text-Davinci-003	11.53	8.95	13.42	15.33	13.28	10.20	16.50	17.75
ChatGPT	12.72	9.33	16.47	17.31	17.83	15.23	20.25	21.76
GPT-4	12.87	10.09	15.96	16.54	24.26	20.85	27.44	29.36
Description-based Frame Induction								
Llama-2 (7B)	5.53	3.31	7.46	9.11	11.59	9.23	13.84	14.66
ChatGLM (6B)	3.17	2.40	3.73	4.56	9.06	7.95	10.02	10.86
GLM (130B)	5.19	4.22	6.05	6.71	15.43	12.68	17.40	20.13
Text-Davinci-003	11.88	10.27	13.34	14.17	18.15	15.41	21.38	21.62
ChatGPT	15.92	12.72	19.00	20.84	19.56	16.83	22.12	23.53
GPT-4	15.75	14.08	16.92	18.83	**28.51**	**25.48**	**31.37**	**32.61**

Table 2: Results of Frame Induction.

Method	**Zero-shot(%)**		Few-shot(%)	
	Ma-F1	Mi-F1	Ma-F1	Mi-F1
Llama-2 (7B)	2.84	2.43	12.32	13.09
ChatGLM (6B)	2.25	1.30	22.93	22.54
GLM (130B)	3.64	2.59	30.93	34.60
Text-Davinci-003	6.23	5.73	22.15	23.45
ChatGPT	13.54	13.95	30.96	33.42
GPT-4	16.91	17.62	38.07	**41.73**

Table 3: Results of Frame Element Induction.

to induce more specific and meaningful FEs such as "protagonist" and "salient_entity".

Frame elements generated by LLMs are incomplete yet redundant. As shown in Figure 3, LLMs may generate incomplete FEs for the `Commerce_buy`, missing crucial FEs such as "money" and "explanation" (in orange). Additionally, they may include redundant FEs like "payment_method" (in red), which duplicates the meaning of "means" regarding the transaction method.

4.5. Results of Frame Relation Induction

LLMs have a limited understanding of weakly associated relations. As shown in Table 5, the F1-score of "Inheritance" relation is 47.31% for GPT-4, whereas the "Using" relation lags behind at a mere 2.48% F1-score. The reason is that "Inheritance" relation represents a strong association between frames, where all FEs in the parent frame have corresponding elements in the child frame. On the other hand, the "Using" relation is a weak form of inheritance, as only some of the FEs in the parent frame have corresponding elements in the child frame.

LLMs have difficulties in identifying fine-grained distinctions among frame relations. For instance, when given the frames `Waking_up` and `Being_awake`, LLMs often incorrectly predict "Inchoative_of" instead of "Precedes" [4]. LLMs struggle to accurately distinguish these fine-grained distinctions among frame relations.

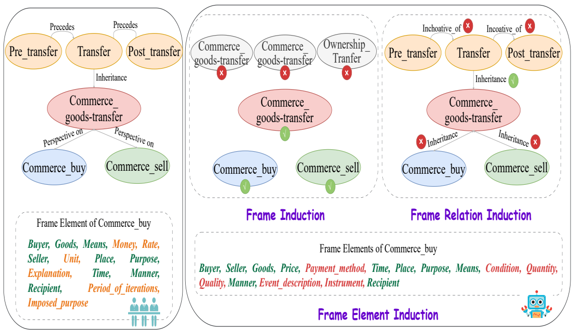

Figure 3: Case Study.

4.6. Case Study

Figure 3 illustrates the FrameNet induction of LLMs, focusing on the `Commerce_buy` frame.

(1) Frame induction: LLMs exhibit limitations in achieving the desired level of abstraction, particularly for higher-level frames like `Transfer`, which is incorrectly predicted as `Commerce_goods-transfer`. This is because `Commerce_goods-transfer` is more concrete than `Transfer`.

(2) Frame element induction: LLMs suffer from issues of incompleteness and redundancy. Crucial elements (in orange) are missed, while redundant elements (in red) are presented.

(3) Frame relation induction: LLMs struggle to differentiate fine-grained distinctions among frame

[4] "Precedes" indicates the temporal or sequential order of events, signifying a sequential relation. On the other hand, "Inchoative_of" implies the beginning or initiation of an action, indicating a state transition.

Method	Zero-shot(%)			Few-shot(%)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Llama-2 (7B)	3.06	2.59	2.85	9.17	13.59	7.83
ChatGLM (6B)	3.44	6.91	3.63	12.47	12.43	12.45
GLM (130B)	14.15	22.31	17.31	21.53	24.82	23.05
Text-Davinci-003	7.80	10.28	8.87	21.99	23.47	22.70
ChatGPT	16.73	19.32	17.93	27.71	20.80	23.76
GPT-4	17.22	22.88	19.65	28.57	24.39	**26.32**

Table 4: Results of Frame Relation Induction.

Relation	ChatGPT			GPT-4		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Inheritance	59.68	9.87	16.94	52.75	42.89	47.31
Subframe	26.47	20.61	23.17	35.29	9.16	14.54
Precedes	66.66	2.24	4.34	87.50	7.86	14.43
Causative_of	22.56	50.00	31.09	39.50	53.33	45.39
Inchoative_of	9.28	68.42	16.35	10.21	73.68	17.94
Perspective_on	11.52	19.84	14.57	32.14	7.08	11.61
Using	32.70	37.29	34.84	77.77	1.26	2.48
See_also	10.34	3.49	5.22	5.28	40.47	9.35
ALL	27.71	20.80	23.76	28.57	24.39	**26.32**

Table 5: Results of Different Frame Relation Induction.

relations. For example, they erroneously predict "Inchoative_of" instead of "Precedes" for the relation between `Pre_transfer` and `Transfer`.

5. Conclusion

In this paper, we construct a Frame-based Conceptual Structure Induction dataset NutFrame. We use prompts to induce conceptual knowledge with LLMs. Extensive experiments indicate that FrameNet induction remains a challenge for existing LLMs. We also provide detailed observations, such as limitations in general frame induction, issues of complete frame element induction, and difficulty in distinguishing subtle frame relations. We hope that our benchmark and findings will facilitate further research on conceptual structure knowledge induction.

Acknowledgements

We thank anonymous reviewers for their insightful comments. This work is supported by the National Key Research and Development Program of China (No.2022ZD0160503), the National Natural Science Foundation of China (No.62176257), and the Youth Innovation Promotion Association CAS. This work is supported by the Four "Batches" Innovation Project of Invigorating Medical through Science and Technology of Shanxi Province (No.2022XM01) and the Science and Technology Cooperation and Exchange Special Project of Shanxi Province (No.202204041101016) .

6. Bibliographical References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL*, pages 86–90.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Jessica A Collins and Ingrid R Olson. 2014. Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic bulletin & review*, 21:843–860.

Robert Alain de Beaugrande and Wolfgang Ulrich Dressler. 1986. *Introduction to text linguistics*. Longman linguistics library. Longman.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 320–335.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Yong Guan, Jiaoyan Chen, Freddy Lecue, Jeff Pan, Juanzi Li, and Ru Li. 2023. Trigger-argument based explanation for event detection. In *Findings of the Association for Computational Linguistics, ACL*, pages 5046–5058.

Shaoru Guo, Chenhao Wang, Yubo Chen, Kang Liu, Ru Li, and Jun Zhao. 2023. EventOA: An event ontology alignment benchmark based on FrameNet and Wikidata. In *Findings of the Association for Computational Linguistics, ACL*, pages 10038–10052.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1725–1744.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 1073–1094.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2302.04023*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Karalyn Patterson, Peter J Nestor, and Timothy T Rogers. 2007. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. COPEN: probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 5015–5035.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 2463–2473.

Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. proscript: Partially ordered scripts generation. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2138–2149.

Stephen W. Smoliar. 1987. J. f. sowa, conceptual structures: Information processing in mind and machine. *Artif. Intell.*, 33(2):259–266.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiwang Yang, Harald Steck, Yang Guo, and Yong Liu. 2012. On top-k recommendation using social networks. In *Sixth ACM Conference on Recommender Systems, RecSys '12*, pages 67–74.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR*.