KnowComp Submission for WMT23 Word-Level AutoCompletion Task

Yi Wu[1,2], Haochen Shi[2], Weiqi Wang[2], Yangqiu Song[2]
[1]University of Wisconsin-Madison, WI, USA
[2]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
ywu676@wisc.edu

Abstract

The NLP community has recently witnessed the success of Large Language Models (LLMs) across various Natural Language Processing (NLP) tasks. However, the potential of LLMs for word-level auto-completion in a multilingual context has not been thoroughly explored yet. To address this gap and benchmark the performance of LLMs, we propose an LLM-based system for the WMT23 Word-Level Auto-Completion (WLAC) task. Our system utilizes ChatGPT to represent LLMs and evaluates its performance in three translation directions: Chinese-English, German-English, and English-German. We also study the task under zero-shot and few-shot settings to assess the potential benefits of incorporating exemplars from the training set in guiding the LLM to perform the task. The results of our experiments show that, on average, our system attains a 29.8% accuracy on the test set. Further analyses reveal that LLMs struggle with WLAC in the zero-shot setting, but performance significantly improves with the help of additional exemplars, though some common errors still appear frequently. These findings have important implications for incorporating LLMs into computer-aided translation systems, as they can potentially enhance the quality of translations. Our codes for evaluation are available at https://github.com/ethanyiwu/WLAC.

1 Introduction

Recent advancements in machine translation, especially due to the development of transformers and pre-trained language models, have been significant (Kong and Fan, 2021; Sun et al., 2023; Mohammadshahi et al., 2022). These methods have yielded impressive results in traditional sentence-level translation tasks (Bahdanau et al., 2015). However, challenges still exist that hinder the further progress of Computer-Aided Translation (CAT) (Esplà-Gomis et al., 2022) systems. Among various components that constitute CAT, Word-level AutoCompletion (WLAC) stands out as a core function (Li et al., 2021; Casacuberta et al., 2022). As shown in Figure 1, WLAC aims to suggest the correct word translation in the target language based on a sequence of human-typed characters and bidirectional context.

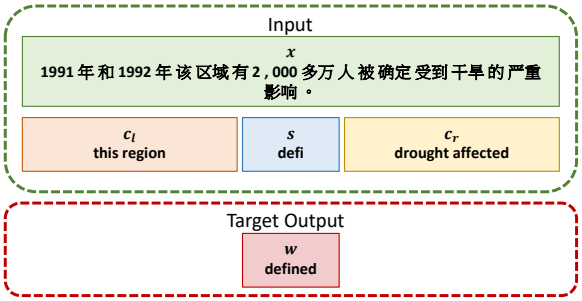

Figure 1: Illustration of the WLAC task for translating from Chinese to English, including various components involved in the translation process. The inputs consist of the source sentence x, the left context in the target sentence c_l, the right context in the target sentence c_r, and the pre-typed character sequence of the word to be predicted s. The task aims to predict the target output word w accurately.

While this task might seem straightforward for seasoned human translators, existing deep-learning approaches struggle to handle it effectively. This can be attributed to the fact that performant translation methods, which rely on pre-trained language models, cannot effectively interpret the typed sequence of characters as they are pre-trained at the token level. Other related studies have either only considered the source contextualization (Huang et al., 2015) or have been unable to handle multilingual translations effectively (Huang et al., 2018). Previous works have demonstrated that transformer-based frameworks, when trained with carefully designed masking or context transformation strategies, can efficiently tackle this task (Yang et al., 2022a,b; Navarro et al., 2022). Yet, these frameworks require extensive training, and their

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 882–889
December 6–7, 2023.

ability to transfer across different languages remains questionable.

With the recent progress made by Large Language Models (LLMs), such as ChatGPT (OpenAI, 2022), researchers have conducted extensive studies regarding their performances on various NLP tasks (Laskar et al., 2023; Liu et al., 2023; Li et al., 2023a; Yu et al., 2023). These LLMs possess advantages, such as ease of deployment and robust multilingual reasoning ability, making them particularly suitable for tasks like WLAC. However, a detailed study of LLMs' application in this context is lacking, and the systematic use of LLMs to assist CAT, especially in performing WLAC, remains unexplored.

This paper addresses these gaps by introducing a system that employs LLMs, specifically ChatGPT, for the WLAC task. ChatGPT has been chosen for its exceptional performance in general natural language tasks, negating the need for deploying or fine-tuning other models. Our approach uses a prompt-engineering-based method to convert all WLAC task inputs into comprehensible natural language sentences. Following this, we synthesize exemplars from the existing training set to facilitate in-context learning with ChatGPT (Wei et al., 2022b). The generations are subsequently parsed using carefully crafted rules to yield the final "predictions" from ChatGPT (Section 3). We then conduct comprehensive experiments using our system, covering scenarios from zero-shot to five-shots. Our experimental results indicate that our system achieves an average accuracy of 29.8% on the testing set. Through error analysis and case studies, we found that LLMs face challenges with WLAC in the zero-shot setting and identified four common types of mistakes that can be particularly addressed in the future. However, ChatGPT's performance significantly improves when provided with additional exemplars, highlighting the crucial role of in-context learning in tackling WLAC for LLMs (Section 4). We will make all codes publicly available upon acceptance of this paper.

2 Preliminaries

2.1 Task Definition

Formally, the objective of the WLAC task (Li et al., 2021) is to predict the target word w using three parts of inputs, which are denoted as the source sequence x, the human-typed characters s, and the translation context c, where $c = (c_l, c_r)$. The translation context consists of left context c_l and right context c_r, where c_l is a sub-sequence of the translated context on the left side of s, and c_r is a subsequence of the translated context on the right side of s. A running example is shown in Figure 1.

	Training	Validation	Test
zh-en	39,473	29,051	16,386
de-en	40,000	29,596	14,564
en-de	40,000	29,895	14,539

Table 1: Statistics regarding the number of data across three translation languages in each split.

Specifically, a notable challenge in training a masked language model for the WLAC task is the incomplete nature of the left and right contexts. These contexts may not necessarily constitute complete sentences; they can consist of partial words or even be empty. As a result, the context c and the typed sequence s do not necessarily provide a fully translated result of the source sequence. Moreover, the training data for this task does not include the complete translated result as a reference. This lack of complete supervision further complicates the establishment of robust training signals for masked language modeling (Li et al., 2023b), especially when compared to traditional translation tasks (Navarro et al., 2022).

2.2 Large Language Models

The emergence of large language models (LLMs) has recently gained the spotlight in the NLP community. GPT3.5 (Brown et al., 2020; Ouyang et al., 2022), ChatGPT (OpenAI, 2022), and LLaMA (Touvron et al., 2023) are some of the notable LLMs that have been well-developed, each boasting an exceptionally vast number of parameters. These LLMs are trained on massive corpora using advanced techniques, such as instruction tuning (Wei et al., 2022a) and reinforcement learning from human feedback (Christiano et al., 2017), on large computational infrastructures. As a result, recent studies have shown that LLMs excel at various downstream tasks, including causal reasoning and grounding (Chan et al., 2023; Wang et al., 2023c; Ou et al., 2023), commonsense reasoning (Fang et al., 2023, 2021b,a; Bian et al., 2023; Wang et al., 2023b), question-answering (Wang et al., 2023a; Qin et al., 2023), translation (Peng et al., 2023; Lu et al., 2023), and data mining tasks (Jin et al., 2023a,b,c). Since the WLAC task demands substantial reasoning and generation capabilities that

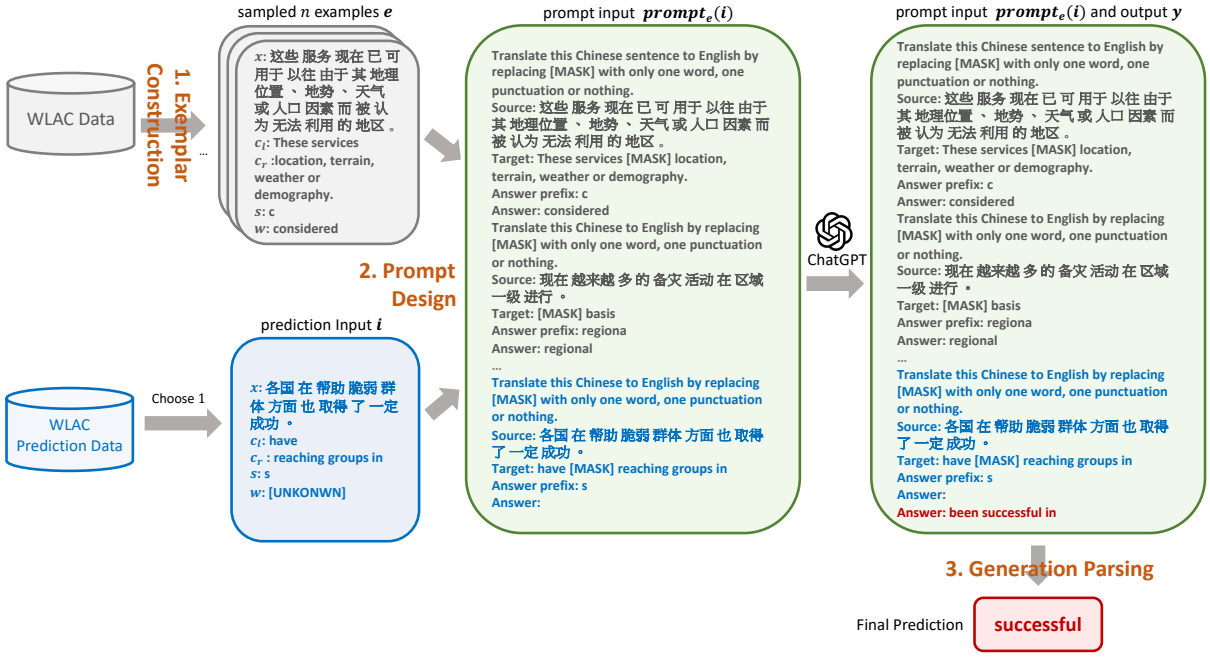

Figure 2: An overview of our framework when dealing with WLAC in translating from Chinese to English (zh-en). The input sentences (marked in blue) are concatenated with pre-constructed exemplars (marked in gray) to form a unified prompt. A large language model (ChatGPT) is then deployed to generate the response (marked in red), which is subsequently parsed to obtain the final prediction y.

rely on bidirectional contexts to accurately predict the target word, LLMs make for an ideal choice to perform WLAC due to their exceptional natural language understanding abilities and ease of deployment.

2.3 Dataset

We use the dataset provided by (Casacuberta et al., 2022) as our primary evaluation benchmark. We select three translation language pairs from the dataset: Chinese to English (zh-en), German to English (de-en), and English to German (en-de). To maintain consistency, we follow the *trn/dev/test* split released in the original dataset. Detailed statistics on the number of data are shown in Table 1.

3 Method

Figure 2 shows an overview of our framework, which consists of three steps: exemplar sampling, prompt design, and generation parsing.

3.1 Exemplar Sampling and Construction

To generate the input prompt for the LLM, we begin by employing random sampling to choose k data instances from the training split of the dataset. These selected instances serve as in-context learning exemplars. In our experiments, we explore different values of $k \in \{0, 1, 5\}$ to evaluate the performance of the LLM in both zero-shot and few-shot scenarios. The aim is to improve the model's familiarity with the task and its capacity to deliver precise answers by incorporating the provided exemplars into its learning process.

3.2 Prompt Design

We then design a natural language prompt to systematically combine both sampled exemplars and every testing data entry from the testing split, which serves as the input for the LLM. To assist the LLM in distinguishing different components of the input and the desired output, we introduce instructive tokens such as "Source" (the source sentence to be translated), "Target" (the target sentence with context c_l and c_r provided), "Answer prefix" (the pre-typed sequence s indicating the target word to be predicted at `[MASK]`), and "Answer" (representing the target prediction word w). For each testing data entry, we construct such a prompted sentence with the "Answer" for the testing entry left blank, awaiting completion. Combining all these prompted sentences creates a comprehensive paragraph of sentence input, as illustrated in Figure 2. This transforms the task into a blank-filling exercise, wherein the model fills in the missing word, the last word in our case.

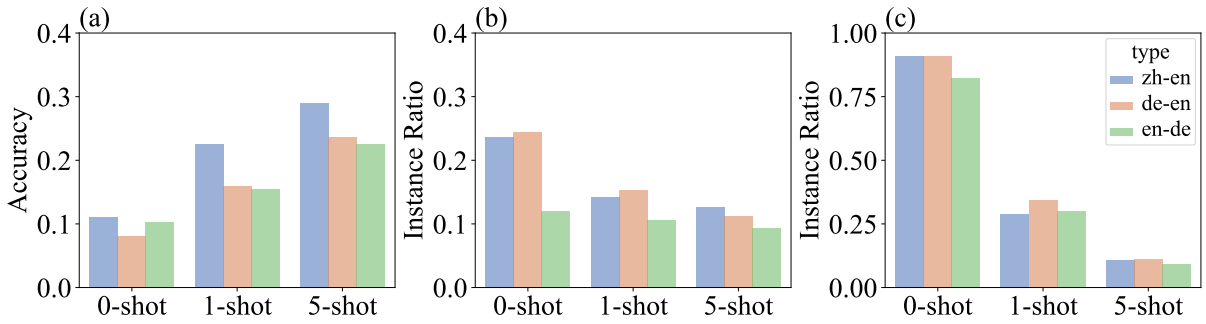

Figure 3: The fault rate and accuracy under different settings. (a) refers to the accuracy. (b) refers to the proportion of instances that don't start with the typed sequence s. (c) refers to the ratio of the generations which contain a sequence of words.

3.3 Generation Parsing

Since the generated content is in free-text form, it requires parsing to identify the predicted word as the final output of the WLAC task. The first step is to remove the cue word, such as "Answer," and separate the remaining sequence of tokens into individual words by splitting at blank spaces. To ensure that the selected word satisfies the constraint of the pre-typed sequence, we search for the first word that starts with the pre-typed sequence s as the final prediction. If such a word does not exist, the first word in the sequence is chosen as the final prediction. For example, if the generated content is "Answer: Hello World" and the pre-typed sequence is "Wo," the word "World" is the final prediction after parsing.

4 Experiments

4.1 Setup

We utilize the official ChatGPT API[1] to access the large language model for sentence completion. The model code employed is `gpt-3.5-turbo`, and the access date is July 2023. The temperature is set to 0.7 when generating to ensure consistent generation, while other control parameters are set to their default values.

4.2 Results

Our experimental results are presented in Table 2. We observe that five-shot prompting yielded the best performance, while not incorporating any training exemplar led to the worst performance. This outcome is reasonable, considering the model may not clearly understand the task objective and reasoning process. Importantly, incorporating just one additional exemplar had a significantly positive impact. This suggests that ChatGPT can quickly learn from provided exemplars and develop a sufficient understanding of the task. The accuracy improvement trends across the three translation directions are consistent, leading us to conclude that ChatGPT can achieve acceptable performance on the WLAC task with the help of training exemplars and in-context learning. However, even with five-shot prompting, the performance is only around 25% in terms of accuracy, leaving a large space for future improvements. Therefore, leveraging more advanced or meticulously designed prompts should be considered further to enhance ChatGPT's performance on the WLAC task.

Task	Shot #		
	0-shot	1-shot	5-shot
Zh - En	11.15	22.66	29.21
En - De	10.26	15.45	22.51
De - En	8.12	15.99	23.66

Table 2: Evaluation results (Accuracy %) on the testing sets of three translation directions.

4.3 Error Analysis

Upon further analysis of the results, we identify two common types of mistakes where the generated output deviates from the targeted answer. The first type of mistake occurs when the generated output fails to begin with the specified sequence s, while the second type of mistake involves the presence of multiple words in the generated output after removing the cue word. As depicted in Figure 3, the overall performance improves significantly as the number of shots increases. Nevertheless, there is a remarkably high rate of faults in generating

[1] https://chat.openai.com/

Source	Context	Generation	Target
CORRECT EXAMPLES			
据报道，受重伤的维和人员得以继续飞行，并与其他机组人员一起成功降落在北基伍省会戈马机场。	It was reported that [MASK]	Answer: seriously	seriously
他要求刚果（金）当局对这起令人发指的袭击事件展开调查，尽快将肇事者绳之以法。	The congo [MASK]	Answer: authorities	authorities
GRAMMATICAL MISTAKE			
她说："我谴责这次袭击，必须以最坚定的态度起诉犯罪者。"	attack and the perpetrators must be prosecuted with the utmost [MASK]	Answer: firmly	firmness
报告说，迫使巴勒斯坦人背井离乡的"胁迫性环境"使巴勒斯坦社会四分五裂，阻碍了自决权的实现。	" coercive environment " that forced the [MASK]	Answer: Palestinian	Palestinians
BOTH CORRECT?			
委员会呼吁联合国大会要求国际法院就占领的法律后果发表紧急咨询意见。	The [MASK]	Answer: committee	called
专家们注意到，欧盟反欺诈办公室就对独立人权组织哈克进行了审查，其结论是："没有发现受怀疑的违规和(或)欺诈行为来影响欧盟的资金"。	experts [MASK]	Answer: noticed	noted
DON'T START WITH TYPED SEQUENCE			
另据报道，在8月18日及21日，以色列国内安全局审问了7个团体中的巴勒斯坦妇女委员会联盟、独立人权组织哈克和保卫儿童 - 巴勒斯坦组织的负责人，据称还对这三人加以威胁。	It was also reported that , on 18 and 21 [MASK]	Answer: August	Actions
谭德塞说："自那时以来，世卫组织已报告了3200多例猴痘确诊病例和一例死亡，这些病例来自包括尼日利亚在内的48个国家和5个世卫组织地区。"	, WHO has reported more than 3200 confirmed [MASK]	Answer: cases	regions
GENERATE A SEQUENCE OF WORDS			
委员会成员克里斯·西多蒂（Chris Sidoti）表示，以色列政府的行动构成了一种非法占领和吞并制度，必须加以解决。	Chris Sidoti said the [MASK]	Answer: Israeli government	Israeli
委员会的成员不是联合国工作人员，他们的工作没有报酬。	The members of the Committee [MASK]	Answer: are not UN staff, their work is voluntary.	not

Table 3: Case studies of generations from ChatGPT. We select generation results from the 5-shot scenario.

word sequences that violate the instruction of using only a single word in the zero-shot approach. The transition from zero-shot to one-shot learning results in a considerable reduction in both fault rates. This indicates that the language model adheres to instructions more accurately by adding a single example. Moreover, the fault rate also further decreases in the five-shot setting.

4.4 Case Studies

To further demonstrate the difficulty of the task and the performance of ChatGPT, we select some generation results in the Chinese-English translation split, as shown in Table 3. Among these cases, we observe four types of tricky but common mistakes. Firstly, although the generated output and the target share the same semantic meaning, the generation is syntactically incorrect. We exemplify two generations that contain grammatical mistakes. Secondly, our approach may generate semantically accurate output that deviates from the target, particularly in cases where the input lacks detailed context. Moreover, the model may generate the content immediately after the context instead of following instructions to find a semantically correct word that matches the typed sequence. Finally, the model does not always comply with the instructions that require it to generate only one word. It may give a phrase or part of a sentence to combine the context into a complete sentence.

4.5 Discussions

While our system achieves acceptable performance, it falls significantly short of the performance achieved by systems last year (Casacuberta et al., 2022). This suggests that additional efforts are required to enhance ChatGPT's performance on the WLAC task, which might include: (a) Incorporating more training exemplars. For instance, increasing the number of training shots to ten or even more could be beneficial. (b) Reframing the exemplar selection problem as a subset selection problem. This approach involves selecting training exemplars based on their similarity to the testing entry or their diversity in relation to other exemplars, as proposed by Ye et al. (2023). (c) Improving the prompt to better leverage both left and right contexts. Additionally, advanced prompting techniques like chain-of-thought (Wei et al., 2022b) could be explored. (d) Incorporating external knowledge for reasoning, such as complex knowledge (Bai et al., 2023), conceptualization (He et al., 2022), and graph reasoning (Liu and Song, 2022; Liu et al., 2022, 2020).

5 Conclusions

In conclusion, this paper presents a novel LLM-prompting system to address the WLAC task. Our findings demonstrate that LLMs are highly capable

problem solvers and adept at learning in context for this particular task, albeit with performance that falls short of previous supervised learning systems. A detailed analysis uncovers several error types that contribute to the limited performance of ChatGPT. Therefore, we urge researchers to devote additional attention to the WLAC task using LLMs.

Acknowledgements

The authors would like to thank the committee of WMT2023, the organizers of the WLAC task, and the anonymous reviewers. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023. Complex query answering on eventuality knowledge graph with implicit logical constraints. *CoRR*, abs/2305.19068.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *CoRR*, abs/2303.16421.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Lemao Liu, Shuming Shi, Taro Watanabe, and Chengqing Zong. 2022. Findings of the word-level autocompletion shared task in WMT 2022. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 812–820. Association for Computational Linguistics.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2022. Cross-lingual neural fuzzy matching for exploiting target-language monolingual corpora in computer-aided translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7532–7543. Association for Computational Linguistics.

Tianqing Fang, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. CKBP v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

Guoping Huang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2015. A new input method for human translators: Integrating machine translation effectively and imperceptibly. In *Proceedings of the Twenty-Fourth*

International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, pages 1163–1169. AAAI Press.

Yafang Huang, Zuchao Li, Zhuosheng Zhang, and Hai Zhao. 2018. Moon IME: neural-based chinese pinyin aided input method with customizable association. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 140–145. Association for Computational Linguistics.

Yiqiao Jin, Yunsheng Bai, Yanqiao Zhu, Yizhou Sun, and Wei Wang. 2023a. Code recommendation for open source software developers. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 1324–1333. ACM.

Yiqiao Jin, Yeon-Chang Lee, Kartik Sharma, Meng Ye, Karan Sikka, Ajay Divakaran, and Srijan Kumar. 2023b. Predicting information pathways across online communities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1044–1056. ACM.

Yiqiao Jin, Xiting Wang, Yaru Hao, Yizhou Sun, and Xing Xie. 2023c. Prototypical fine-tuning: Towards robust performance under varying data sizes. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12968–12976. AAAI Press.

Yawei Kong and Kai Fan. 2021. Probing multi-modal machine translation with pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3689–3699. Association for Computational Linguistics.

Md. Tahmid Rahman Laskar, M. Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy X. Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 431–469. Association for Computational Linguistics.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *CoRR*, abs/2304.05197.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023b. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14022–14040. Association for Computational Linguistics.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. GWLAN: general word-level autocompletion for computer-aided translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4792–4802. Association for Computational Linguistics.

Xin Liu, Jiayang Cheng, Yangqiu Song, and Xin Jiang. 2022. Boosting graph structure learning with dummy nodes. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 13704–13716. PMLR.

Xin Liu, Haojie Pan, Mutian He, Yangqiu Song, Xin Jiang, and Lifeng Shang. 2020. Neural subgraph isomorphism counting. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1959–1969. ACM.

Xin Liu and Yangqiu Song. 2022. Graph convolutional networks with dual message passing for subgraph isomorphism counting and matching. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 7594–7602. AAAI Press.

Xin Liu, Yuan Tan, Zhenghang Xiao, Jianwei Zhuge, and Rui Zhou. 2023. Not the end of story: An evaluation of chatgpt-driven vulnerability description mappings. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3724–3731. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *CoRR*, abs/2303.13809.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8348–8359. Association for Computational Linguistics.

Ángel Navarro, Miguel Domingo, and Francisco Casacuberta. 2022. Prhlt's submission to WLAC 2022. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages

1182–1186. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Jiefu Ou, Adithya Pratapa, Rishubh Gupta, and Teruko Mitamura. 2023. Hierarchical event grounding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13437–13445. AAAI Press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *CoRR*, abs/2303.13780.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

Simeng Sun, Maha Elbayad, Anna Sun, and James Cross. 2023. Efficiently upgrading multilingual machine translation models to support more languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1505–1519. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. *CoRR*, abs/2305.14869.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023c. COLA: contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5253–5271. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Cheng Yang, Siheng Li, Chufan Shi, and Yujiu Yang. 2022a. IIGROUP submissions for WMT22 word-level autocompletion task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1187–1191. Association for Computational Linguistics.

Hao Yang, Hengchao Shang, Zongyao Li, Daimeng Wei, Xianghui He, Xiaoyu Chen, Zhengzhe Yu, Jiaxin Guo, Jinlong Yang, Shaojun Li, Yuanchang Luo, Yuhao Xie, Lizhi Lei, and Ying Qin. 2022b. Hw-tsc's submissions to the WMT22 word-level auto completion task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1192–1197. Association for Computational Linguistics.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.