TRANSAGENTS: Build Your Translation Company with Language Agents

Minghao Wu[1] **Jiahao Xu**[2,3] **Longyue Wang**[3]*
[1]Monash University [2]Nanyang Technological University [3]Tencent AI Lab
minghao.wu@monash.edu jiahao004@e.ntu.edu.sg vinnylywang@tencent.com

Abstract

Multi-agent systems empowered by large language models (LLMs) have demonstrated remarkable capabilities in a wide range of downstream applications. In this work, we introduce TRANSAGENTS, a novel multi-agent translation system inspired by human translation companies. TRANSAGENTS employs specialized agents — Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader — to collaboratively produce translations that are accurate, culturally sensitive, and of high quality. Our system is *flexible*, allowing users to configure their translation company based on specific needs, and *universal*, with empirical evidence showing superior performance across various domains compared to state-of-the-art methods. Additionally, TRANSAGENTS features a *user-friendly* interface and offers translations at a cost approximately $80\times$ *cheaper* than professional human translation services. Evaluations on literary, legal, and financial test sets demonstrate that TRANSAGENTS produces translations preferred by human evaluators, even surpassing human-written references in literary contexts. Our live demo website is available at https://www.transagents.ai/. Our demonstration video is available at https://www.youtube.com/watch?v=p7jIAtF-WKc.

1 Introduction

Large language models (LLMs) have revolutionized the field of natural language processing and artificial intelligence, achieving remarkable progress in various downstream applications (Ouyang et al., 2022; Sanh et al., 2022; OpenAI, 2023; Anil et al., 2023b; Touvron et al., 2023a,b; Anil et al., 2023a; Mesnard et al., 2024; Dubey et al., 2024). The superior capabilities of LLMs also empower a wide range of multi-agent systems (Yao et al., 2023; Wang et al., 2023c; Dong et al., 2023), enhancing their efficiency and effectiveness in diverse domains, including software development (Qian et al., 2023; Hong et al., 2023), simulation (Park et al., 2022, 2023; Li et al., 2023), gaming (Xu et al., 2023b), and more.

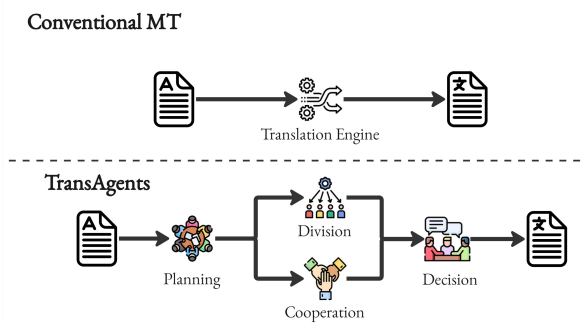

Figure 1: Compared to conventional machine translation (MT) systems that utilize a single MT engine, TRANSAGENTS leverages the collaboration among multiple language agents, each powered by large language models (LLMs), for translation.

Among all the above, one particularly exciting application of multi-agent systems is in the field of machine translation (MT). MT systems, which typically rely on a single model to perform the translation, have achieved considerable success (Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017; Costa-jussà et al., 2022). However, these systems often encounter difficulties in accurately handling nuances, context, and idiomatic expressions (Freitag et al., 2021; Thai et al., 2022). This limitation highlights the need for a superior approach that can handle the subtleties of human language more effectively.

Consequently, to address the aforementioned limitations of recent MT systems, we draw inspiration from the traditional translation industry's workflow and propose TRANSAGENTS as shown in Figure 1. Similar to a human translation company, TRANSAGENTS functions as a virtual multi-agent translation company. It mitigates the challenge of generating high-quality translations by dividing

*Longyue Wang is the corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 131–141

the translation process into several steps and utilizing the collaborative efforts of multiple specialized agents. More specifically, in TRANSAGENTS, each agent is designed to manage specific aspects of the translation process, to produce accurate and natural translations akin to those of human translators. Each of our agents plays a specialized role, including Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader. Together, these agents replicate the traditional human translation process, delivering translations that are accurate, culturally sensitive, and of high quality. Finally, we evaluate TRANSAGENTS alongside other state-of-the-art translation systems using three test sets from the literary, legal, and financial domains. Our experimental results show that, despite lower d-BLEU scores, the translations from TRANSAGENTS are significantly more preferred by human evaluators from the target audience compared to other state-of-the-art translation systems. Notably, the literary translations provided by TRANSAGENTS are even more preferred than the human-written reference translations.

Our system is featured by the following characteristics:

- **Flexible**: TRANSAGENTS allows users to configure their translation company based on their specific needs, such as the number of employees for each role, the source and target languages, and the backbone of language agents.
- **Universal**: Empirical results indicate that TRANSAGENTS significantly outperforms other methods in translations across various domains, according to human evaluations.
- **User-Friendly**: We design a straightforward and intuitive user interface to enhance the user experience as shown in Figure 3. This interface is easy to navigate, allowing users to access the system's functionalities effortlessly.
- **Cost-Effective**: The cost of translating documents using TRANSAGENTS is approximately $80\times$ cheaper than professional translation services as described in Section 4.4.

2 Related Work

Large Language Models Large language models (LLMs) have significantly transformed the field of artificial intelligence. These models are pretrained on extensive text corpora to predict the next word in a sentence, which allows them to understand and generate human-like text (Brown et al., 2020; Chowdhery et al., 2022; Anil et al., 2023b; Touvron et al., 2023a,b; Anil et al., 2023a,a; Yang et al., 2024). After the initial pretraining phase, LLMs undergo supervised fine-tuning (SFT) or instruction tuning (IT). This process helps align the models more closely with human instructions, enhancing their ability to perform specific tasks (Sanh et al., 2022; Chung et al., 2022; Tay et al., 2023; Shen et al., 2023; Wu et al., 2024b). Recent developments in the field include the use of synthetic datasets generated by LLMs for fine-tuning. Additionally, reinforcement learning from human feedback (RLHF) is employed to further improve the models' performance and reliability (Ouyang et al., 2022; Hejna et al., 2023; Ethayarajh et al., 2024; Hong et al., 2024; Meng et al., 2024).

Multi-Agent Systems Intelligent agents are designed to understand their environments, make informed decisions, and respond appropriately (Wooldridge and Jennings, 1995). Recent multi-agent systems utilize collaboration among multiple agents based on LLMs to tackle complex problems or simulate real-world environments effectively (Guo et al., 2024), such as software development (Qian et al., 2023; Hong et al., 2023), multi-robot collaboration (Mandi et al., 2023; Zhang et al., 2023), text generation (Liang et al., 2023), and simulate societal, economic, and gaming environments (Park et al., 2023; Xu et al., 2023b).

Machine Translation Machine translation (MT) has seen remarkable advancements in recent years (Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017; Gu et al., 2018; Fan et al., 2021; Communication et al., 2023). However, these improvements are predominantly at the sentence level. Recent research has shifted focus towards incorporating contextual information to enhance translation quality beyond individual sentences (Wang et al., 2017; Wu et al., 2023; Herold and Ney, 2023; Wu et al., 2024c). This involves leveraging document-level context to provide more accurate translations. Additionally, large language models (LLMs) have demonstrated superior capabilities in MT, further pushing the boundaries of translation quality (Xu et al., 2023a; Robinson et al., 2023; Wang et al., 2023a; Wu et al., 2024a).

Ours In this work, we introduce TRANSAGENTS, a general-purpose multi-agent framework that harnesses collaborative efforts among agents for translation. These language agents are powered by the

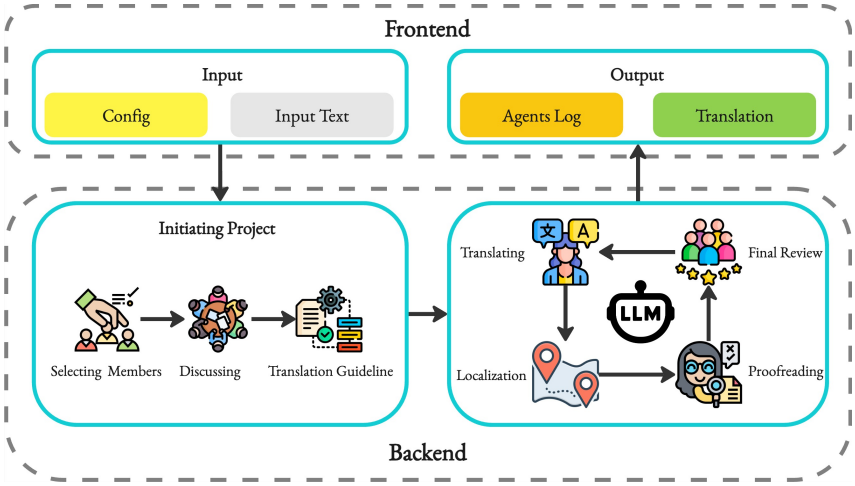

Figure 2: The overview of TRANSAGENTS, including the **Frontend** and **Backend** modules.

latest state-of-the-art LLMs.

3 TRANSAGENTS

Our demo system TRANSAGENTS is implemented as a web application, built using Streamlit.[1] The system comprises two main modules: a front-end and a back-end. As illustrated in Figure 2, the frontend module is responsible for accepting user input, including the document to be processed and task configurations (Section 3.1). The backend module, on the other hand, handles the translation of the given document by orchestrating the collaborative efforts of our language agents (Section 3.2). Additionally, we present a step-by-step walkthrough of TRANSAGENTS in Section 3.3.

3.1 Frontend Design

Task Configuration In addition to accepting documents for translation from users, we also allow users to configure their tasks. As shown in Figure 3, this includes specifying the backbone of the language agents, selecting the source and target languages, determining the number of candidates for various roles in the company, and more.

Progress Visualization As shown in Figure 3, when the language agents collaborate with each other, we visualize *translation progress checkpoints* and *multi-agent conversations* in the user interface, allowing users to monitor the progress of the translation. This feature provides insights into the decision-making process of the agents, making it easier to understand how translations are derived.

[1] https://streamlit.io/

3.2 Backend Design

Agentic Backbone In our system, we allow users to select various large language models as the backbone of their translation tasks. Users can choose from a range of state-of-the-art large language models, including but not limited to GPT-4, GPT-4o, and others. This selection ensures that users can find the most suitable model for their specific translation requirements. This flexibility not only enhances the quality and accuracy of translations but also allows users to experiment and find the perfect balance between speed, precision, and contextual understanding.

Role Playing TRANSAGENTS mirrors the traditional translation pipeline employed by human translation companies, ensuring an effective and efficient workflow. In our system, we assign distinct roles to language agents by defining specific system prompts tailored to their functions, including the Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader. We leverage large language models (LLMs) to create detailed prompts for each role. These prompts guide the language agents, ensuring they understand their specific tasks and responsibilities within the translation pipeline.

Translation Workflow We illustrate the workflow of TRANSAGENTS in Figure 2. Upon receiving the document to be translated and the task configuration from the user, the Senior Editor first selects appropriate agents for the translation task and prepares the translation guidelines in collaboration with the Junior Editor. The Junior Editor adds as much detail as possible to the translation guidelines,

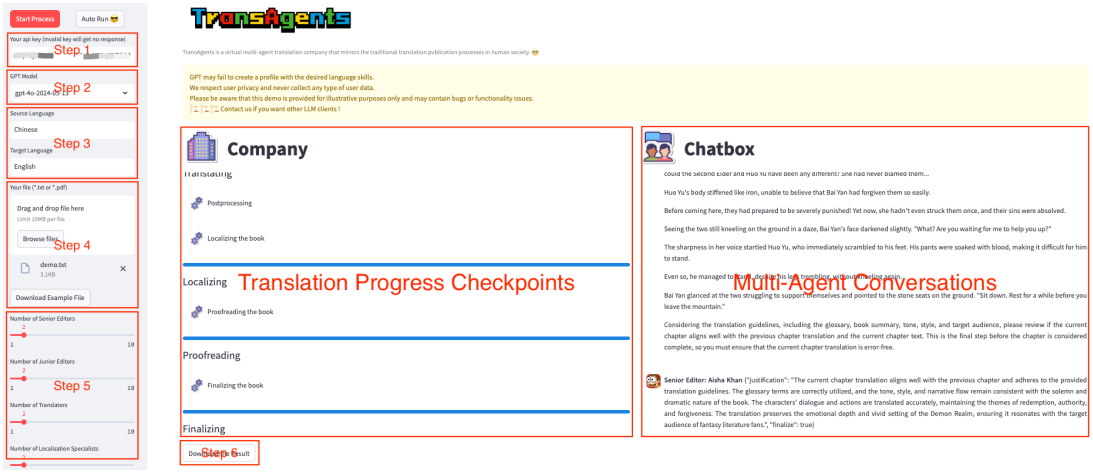

Figure 3: The user interface and step-by-step walkthrough of TRANSAGENTS.

while the Senior Editor is responsible for removing redundant information, refining the guidelines until they are precise and clear. Following this, the Senior Editor and Junior Editor work closely with the Translator, Localization Specialist, and Proofreader. The Junior Editor provides initial feedback on the translations in collaboration with the Translator, Localization Specialist, and Proofreader. The Senior Editor then evaluates whether the translations meet the required quality criteria. Finally, the Senior Editor reviews the quality of the translations. If the translations meet the required standards, they are delivered to the user. Otherwise, they are sent back to the translator for further improvements.

3.3 System Walkthrough

We present a complete walkthrough for using our system in Figure 3:

- **Step 1**: Enter the user's API key;
- **Step 2**: Select the LLM as the backbone of language agents;
- **Step 3**: Specify the source language of the document to be translated and the desired target language for translation;
- **Step 4**: Upload the document to be translated;
- **Step 5**: Set the number of employees for each role in the translation company;
- **Step 6**: Click the start button in the upper right corner to initiate the multi-agent translation process. Once the translation is complete, the user can download the translated document.

4 Experiments

In this section, we first introduce our experimental setup in Section 4.1, followed by presenting the results from both automatic evaluation (Section 4.2) and human evaluation (Section 4.3).

4.1 Setup

Datasets We evaluate our models on three Chinese-English test sets from the literary, legal, and financial domains. The literary test set, sourced from Wang et al. (2023b), comprises 240 chapters from 20 web novels, with each chapter averaging approximately 1,400 words. The legal test set is an in-house collection of 500 contracts, each containing around 68K words. Similarly, the financial test set is an in-house collection of 500 financial reports, with each report containing roughly 83K words. The figures and charts in the financial reports are removed. Both the legal and financial test sets are manually translated by professional translators and reviewed by lawyers and accountants.

Evaluation In this work, we evaluate the translation quality with two methods: automatic evaluation and human evaluation. We leverage d-BLEU (Papineni et al., 2002; Liu et al., 2020) to evaluate the translation quality at the document level, and describe the details for human evaluation in Section 4.3.

Models In this work, the backbone of the language agents in TRANSAGENTS is GPT-4-TURBO.[2] We compare TRANSAGENTS with NLLB-200-3.3B (Costa-jussà et al., 2022), GOOGLE-TRANS,[3] GPT-4-TURBO, and GPT-4O.[4]

[2] gpt-4-turbo-2024-04-09
[3] https://translate.google.com/
[4] gpt-4o-2024-05-13

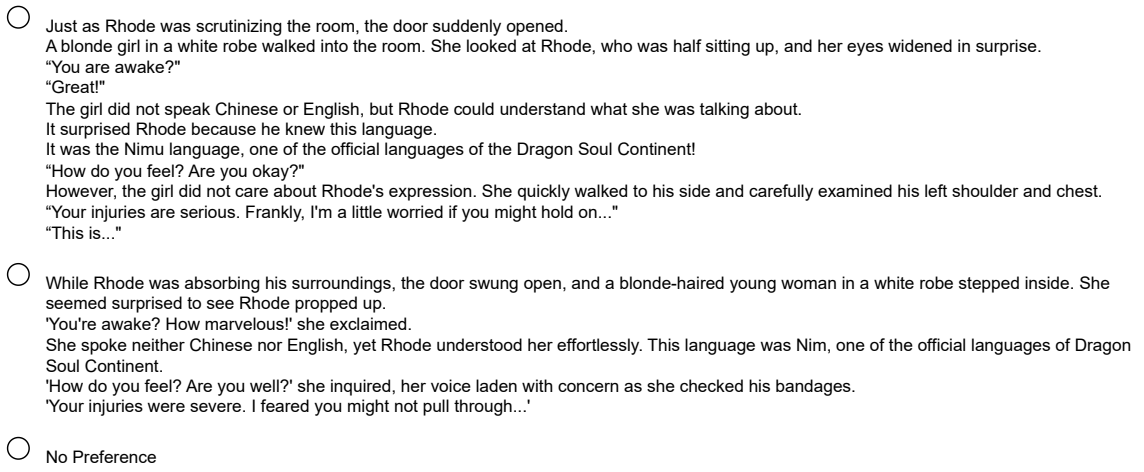

Figure 4: The user interface for human evaluation. The human evaluators select their preferred translations.

	Literary	Legal	Financial
NLLB-200-3.3B	20.2	24.8	26.5
GOOGLETRANS	47.3	37.9	35.8
GPT-4-TURBO	**47.8**	38.9	36.7
GPT-4O	46.8	**39.0**	**37.8**
TRANSAGENTS	25.0	30.9	32.9

Table 1: d-BLEU given by TRANSAGENTS and baselines on three test sets. **The BEST results are highlighted in bold**.

	Literary	Legal	Financial
NLLB-200-3.3B	10.2	15.3	14.8
GOOGLETRANS	38.5	28.9	31.8
GPT-4-TURBO	41.9	30.5	33.9
GPT-4O	43.4	32.7	34.8
TRANSAGENTS	**55.5**	**39.9**	**37.9**

Table 2: Winning rate (WR; %) given by TRANSAGENTS and baselines on three test sets. **The BEST results are highlighted in bold**.

4.2 Automatic Evaluation

We present our results in Table 1. Interestingly, TRANSAGENTS performs poorly in terms of d-BLEU, achieving the lowest scores among all the compared methods. However, these low scores do not necessarily imply poor performance of our approach, as typical references used for calculating d-BLEU scores often exhibit poor diversity and tend to concentrate around translationese language (Freitag et al., 2020). Our results also align with the findings from Thai et al. (2022), where automatic metrics cannot accurately reflect human preference. To confirm this claim, we conduct human evaluation and present the results in Section 4.3.

4.3 Human Evaluation

In this section, we introduce how we conduct human evaluation in this work and present our results.

Setup In the real-world application, it is not necessary for the readers to understand the original language, so we only provide the translated text given by different models and its corresponding reference translation to human evaluators, and require the human evaluators to select their preferred translation. It is hard for human evaluators to ensure the evaluation quality when evaluating the very long documents, so we split the whole document into segments containing approximately 200 English words. For each test set, we employ five human evaluators from the corresponding target audience. For literary test sets, we hire human evaluators from online forum for web novel.[5] Furthermore, we employ the master students majoring in law and finance in U.S. to evaluate the translations. The translation and its reference are anonymized when presented to the human evaluators and their order is randomly shuffled to avoid the potential bias on the position. Due to budget constraints, we only evaluate roughly 500 segments for each test set, and pay $0.5 USD for each annotation. We present the user interface for human evaluation in Figure 4.

Results We present the results in Table 2. TRANSAGENTS significantly outperforms all the baselines in terms of winning rate. Notably, TRANSAGENTS is even more preferred over the human-written reference translations on the literary test set. However, human evaluators still favor the

[5] https://www.reddit.com/r/WebNovels/

Original Text	第834章回归圣地（二）[OMITTED] 第835章回归圣地（三）[OMITTED]
REFERENCE	Chapter 834 Return to the Sacred Land (2) [OMITTED] Chapter 835 Return to the Sacred Land (3)
GPT-4O	Chapter 834: Return to the Holy Land (Part Two) [OMITTED] Chapter 834: Return to the Sacred Land (Part Three)
TRANSAGENTS	Chapter 834: Return to the Sacred Land (Part Two) [OMITTED] Chapter 835: Return to the Sacred Land (Part Three)

Table 3: Case study for translation consistency. The text highlighted in red indicates inconsistent translations across different chapters. The text highlighted in blue indicates consistent translations.

human-written reference translations on the legal and financial test sets. The inter-annotator agreements are 0.64, 0.78, and 0.72 for the literary, legal, and financial test sets, respectively, as measured by Cohen's κ coefficient (Cohen, 1960). These values indicate substantial agreement among the annotators for all three test sets. We believe this discrepancy arises because the evaluation criteria differ across various domains. The readers of literary texts commonly have higher standards for stylistic language and cultural nuances, while the readers of legal and financial documents prioritize precision in language. These findings pave the way for future research.

4.4 Cost Analysis

The American Translators Association advises a baseline fee of $0.12 USD per word for professional translation services,[6] which translates to $168.48 USD per chapter for the literary test set. In contrast, employing TRANSAGENTS for translation purposes incurs a total cost of approximately $500 USD for the entire literary test set, which is equivalent to about $2.08 USD per chapter. Consequently, using TRANSAGENTS for translating literary texts can result in an $80\times$ decrease in translation expenses.

5 Case Study

In this section, we present two case studies from literary test set to demonstrate the superiority of TRANSAGENTS.

[6] https://unbabel.com/translation-pricing-how-does-it-work/

Original Text	慕言君仅仅睡了两个时辰，眼睛就睁开。
REFERENCE	Mu Yanjun only slept for four hours before his eyes opened.
GPT-4O	Mu Yanjun only slept for two hours before his eyes opened.
TRANSAGENTS	After only four hours, Mu Yanjun's eyes opened once more.

Table 4: Case study for culture adaptation. The text highlighted in red indicates incorrect translations. The text highlighted in blue indicates correct translations.

Translation Consistency Ensuring consistency from the beginning to the end of a document is essential. As shown in Table 3, the chapter titles in the original text are consistent, except for the index. While all translation methods deliver semantically accurate results, only REFERENCE and TRANSAGENTS achieve consistency across various chapters. In contrast, GPT-4O has difficulty maintaining this consistency. This highlights that TRANSAGENTS can maintain consistency throughout the entire translation process.

Cultural Adaptation For translation systems to be truly effective, they must incorporate an understanding of cultural and historical contexts. In traditional Chinese timekeeping, a 时辰 ("shichen") is equivalent to two hours in the modern time system. Therefore, 两个时辰 (two "shichen") is equal to four hours. As shown in Table 4, both REFERENCE and TRANSAGENTS correctly translate 两个时辰 to four hours, while GPT-4O fails to convert "shichen" to the modern time system and mistranslates 两个时辰 as two hours. This highlights that TRANSAGENTS has a superior ability to handle culturally specific terms and accurately translate them into the modern context.

6 Conclusion

In this work, we introduce TRANSAGENTS, a novel multi-agent translation system inspired by the traditional human translation process, characterized by its flexibility, universality, user-friendliness, and cost-effectiveness. TRANSAGENTS leverages the collaborative efforts of specialized agents, including a Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader. Our experimental results, derived from test sets across literary, legal, and financial domains, highlight the superior performance of TRANSAGENTS. Although

TRANSAGENTS achieves lower d-BLEU scores compared to other state-of-the-art systems, its translations are significantly more preferred by human evaluators. Our case study also demonstrates the effectiveness of TRANSAGENTS with regard to translation consistency and culture adaptation.

7 Limitations

Translation Latency While TRANSAGENTS is obviously faster than a human translator, it is considerably slower compared to conventional MT systems. This increased latency is due to the extensive communication required among the language agents in TRANSAGENTS.

Evaluation The shortcomings of the BLEU metric are well-documented within the MT literature. Due to budget constraints, our human evaluation covers only a subset of translations. These limitations may impact the reliability of our evaluation.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023a. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023b. Palm 2 technical report. *CoRR*, abs/2305.10403.

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.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Y. Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *CoRR*, abs/2308.11596.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration code generation via chatgpt. *CoRR*, abs/2304.07590.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: model alignment as prospect theoretic optimization. *CoRR*, abs/2402.01306.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *CoRR*, abs/2402.01680.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. 2023. Contrastive preference learning: Learning from human feedback without RL. *CoRR*, abs/2310.13639.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: monolithic preference optimization without reference model. *CoRR*, abs/2403.07691.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *CoRR*, abs/2308.00352.

Nian Li, Chen Gao, Yong Li, and Qingmin Liao. 2023. Large language model-empowered agents for simulating macroeconomic activities. *CoRR*, abs/2310.10436.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *CoRR*, abs/2305.19118.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhao Mandi, Shreeya Jain, and Shuran Song. 2023. Roco: Dialectic multi-robot collaboration with large language models. *CoRR*, abs/2307.04738.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *CoRR*, abs/2405.14734.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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*.

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.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022 - 2 November 2022*, pages 74:1–74:18. ACM.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *CoRR*, abs/2307.07924.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *CoRR*, abs/2305.14705.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world

literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. 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. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023c. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *CoRR*, abs/2302.01560.

Michael J. Wooldridge and Nicholas R. Jennings. 1995. Intelligent agents: theory and practice. *Knowl. Eng. Rev.*, 10(2):115–152.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024b. LaMini-LM: A diverse herd of distilled models from large-scale instructions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–964, St. Julian's, Malta. Association for Computational Linguistics.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024c. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023b. Exploring large language models for communication games: An empirical study on werewolf. *CoRR*, abs/2309.04658.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. *CoRR*, abs/2307.02485.