Curse of bilinguality: Evaluating monolingual and bilingual language models on Chinese linguistic benchmarks

Anonymous ACL submission

Abstract

We investigate cross-lingual transfer in large language models (LLMs) trained on two high-resource languages, English and Chinese. Four monolingual Chinese and four bilingual English–Chinese models are evaluated on two Chinese linguistic benchmarks. Monolingual models consistently outperform the bilingual ones on 12 out of 55 tasks, a result indicating negative transfer from English to Chinese. Additionally, we carry out a feature attribution analysis in a monolingual and a bilingual model, showing that the differences in their performance may be explained by more predictable attribution patterns in the monolingual model. Our findings have implications for the ongoing effort of training bilingual LLMs.

1 Introduction

In multilingual NLP, cross-lingual transfer is traditionally described in positive terms. For example, a model's performance in low-resource languages can be improved by leveraging transfer from high-resource languages. At the same time, adding low-resource languages to the training data may cause a model to perform worse in high-resource languages because of the *negative* cross-lingual transfer, a phenomenon known as the curse of multilinguality (Conneau et al., 2020). Despite the abundance of studies that address this problem (Blevins et al., 2024; Wang et al., 2020; Pfeiffer et al., 2022, etc.), they primarily focus on multilingual LLMs trained on multiple languages with very unbalanced amounts of data per language.

What happens, however, when a model is trained on exactly two high-resource languages? English and (Mandarin) Chinese are the two languages with the largest amounts of data available for training, and the recent years have seen a surge in the development of LLMs for both languages. While a few Chinese models are monolingual (e.g., Sun et al., 2021; Zhang et al., 2021; Zeng et al., 2021), most others are bilingual (i.e., trained on a mix of English and Chinese data: Bai et al., 2023; Yang et al., 2023; Young et al., 2024) or multilingual (see an overview by Huang et al., 2025). While bilingual and multilingual models show high performance on some English benchmarks (e.g., Zeng et al., 2024), it is unclear whether in Chinese linguistic tasks they always outperform their monolingual counterparts.

In this paper, we study cross-lingual transfer effects in bilingual Chinese–English LLMs. We evaluate four monolingual Chinese models and four bilingual Chinese–English models on two commonly used Chinese linguistic benchmarks. For some of the paradigms in these benchmarks, monolingual models (including the relatively small monolingual Chinese BERT) consistently outperform bilingual models, indicating negative transfer from English to Chinese. We then present an interpretability analysis using feature attribution methods on two selected models, showing that the bilingual model may be worse at understanding the relations between words in the target sentences than the monolingual one. We are making our code publicly available in case of acceptance.

2 Method

2.1 Models

We consider a diverse set of pretrained transformer-based LLMs. Whereas many *multilingual* LLMs support both Chinese and English, we focus on the cross-language transfer specifically from English to Chinese. To eliminate possible influences from other languages, we work with bilingual (not multilingual) models. Specifically, we select four monolingual Chinese and four bilingual Chinese–English models based on their performance on common benchmarks and their number of parameters, to cover a variety of model sizes while staying within the limits of our available computational resources. The models and their number of parame-

Model	# param.	Languages
ERNIE	10B	Chinese
CPM	2.6B	Chinese
PANGU	2.6B	Chinese
BERT	0.11B	Chinese
QWEN	14B	Chinese–English
BAICHUAN	7B	Chinese–English
YI	6B	Chinese–English
CHATGLM	6B	Chinese–English

Table 1: Monolingual and bilingual models we consider.

ters are listed in Table 1. Note that the monolingual models (except ERNIE) generally have fewer parameters, giving the bilingual models a possible advantage thanks to their size. In all cases, we use HuggingFace implementations.

The monolingual Chinese models include (1) Ernie-3.0 (Sun et al., 2021), which combines masked and autoregressive training objectives and is trained on 4TB of both textual data and structured knowledge graphs, (2) CPM (Zhang et al., 2021) trained on 100GB Chinese text and tailored to Chinese NLP tasks, (3) Pangu-alpha-2.6B (Zeng et al., 2021), the smallest of the Pangu family of autoregressive models, trained on 100GB of Chinese text, and (4) Chinese BERT (Devlin et al., 2019), a much smaller model considered for reference.

The bilingual Chinese–English models include (1) Qwen (Bai et al., 2023), the base Qwen-family model trained on 3 trillion tokens, (2) Baichuan-7B (Baichuan, 2023), the smaller of the first-generation Baichuan models, trained on 1.2 trillion tokens, (3) Yi-6B (Young et al., 2024), a Yi-family model trained on a 3.1 trillion high-quality Chinese–English tokens, and (4) ChatGLM3-6B (Zeng et al., 2024), a GLM series model optimized for Chinese question answering and dialogue.

2.2 Benchmarks

We evaluate our models on two commonly used linguistic benchmarks of minimal pairs in Chinese: CLiMP (Xiang et al., 2021) and SLING (Song et al., 2022). CLiMP is the Chinese adaptation of the English BLiMP benchmark (Warstadt et al., 2020). It has been criticized for its use of translations that do not naturally reflect Chinese linguistic phenomena. To address this limitation, the second benchmark, SLING, derives its minimal pairs from naturally occurring annotated Chinese sentences and applies syntactic and lexical transformations specifically designed for Chinese grammar, offering a more ecologically valid evaluation framework. Together, these benchmarks contain 18 Chinese linguistic phenomena sub-divided into 55 paradigms with more than 50k minimal pairs of sentences.

For most of the paradigms in the evaluation benchmarks, each minimal pair consists of one grammatical and one ungrammatical sentence. For example, in the SLING *Alternative Question* paradigm, the sentence with the 吗 *(ma)* particle is always ungrammatical, since this particle can only be used in yes–no (but not alternative) questions:

(1) 她们 是 飞行员 还是 制片人 [吗*] ?
they be pilot or producer [Q*] ?
'Are they pilots or producers?'

However, in eight SLING *Anaphor* paradigms (*baseline female/male, baseline cl female/male, baseline cl man female/male, baseline cl men female/male*), both sentences are grammatical. A model's score in these paradigms indicates its preference towards one or the other sentence (i.e., bias) rather than accuracy, e.g., for *baseline female*:

(2) 女队员 攻击了 [她 / 他] 。
female.team.member attacked [she / he] .
'The female team member attacked her/him.'

2.3 Evaluation

We use the standard method of evaluating the models on minimal pairs. In each pair, sentence perplexity (or pseudo-perplexity, for masked models) values are computed, and the sentence with a lower perplexity is taken to reflect the model's preference. This preference is then compared to the ground-truth data, and the model's accuracy for each paradigm is computed.

For each paradigm, we then compare the accuracy values of the 4 monolingual models against those of the 4 bilingual models. In case of positive cross-lingual transfer, one could expect the bilingual models to show higher accuracy values. However, if we observe that for some of the paradigms the monolingual models (which are also generally smaller) consistently outperform the bilingual ones, this can be seen as evidence of negative cross-lingual transfer.

The evaluations and analyses were run on a single Nvidia V100 GPU with 32GB memory, with a total time of 30 hours. We provide the results below, followed by a feature attribution analysis.

	Monolingual models				Bilingual models			
Paradigm	ERNIE	CPM	PANGU	BERT	QWEN	BAICHUAN	YI	CHATGLM
Coverb								
—”— with	82.3	61.7	73.5	84.7	**86.2**	**84.9**	**84.8**	**84.8**
Verb complement								
—”— res adj	59.7	25.9	59.3	87.6	**92.1**	**95.2**	**91.1**	**90.9**
—”— res verb	**92.8**	**96.7**	**90.1**	**96.2**	61.2	65.7	64.2	61.4
Alternative Question								
haishi ma	**94.6**	**85.8**	10.0	**93.1**	9.8	26.6	6.5	**64.0**
Anaphor (Gender)								
baseline female	**92.9**	**89.8**	**95.9**	**86.7**	32.1	66.2	70.3	67.1
Anaphor (Number)								
baseline cl female	**99.5**	**77.9**	0.0	**99.4**	10.1	16.2	29.4	**40.7**
baseline cl male	**99.9**	**75.1**	0.0	**99.6**	26.0	42.9	**47.6**	45.3
baseline cl men female	**99.5**	**88.8**	0.0	**99.4**	5.9	9.7	25.3	**34.8**
baseline cl men male	**100**	**87.6**	0.0	**100**	17.9	38.0	38.9	**43.2**
baseline men female	**99.3**	**51.8**	0.0	**98.0**	6.7	9.4	28.7	**41.4**
cl men self female	**98.3**	**96.2**	0.0	**100**	87.5	**95.4**	84.0	77.9
cl self female	**99.2**	**88.8**	0.0	**99.9**	74.8	**82.8**	62.4	70.2
Definiteness Effect								
every	**96.2**	**92.5**	87.7	**94.6**	**88.0**	69.2	58.7	84.9
Polarity Item								
even wh	85.8	42.3	47.7	52.4	**97.7**	**98.4**	**96.9**	**98.0**
more or less	**98.3**	**98.6**	**97.6**	**97.9**	86.2	96.8	93.3	79.5
Relative Clause								
rc resumptive pronoun	54.8	18.6	11.8	42.7	**64.3**	**77.8**	**68.1**	**60.8**

Table 2: The models' performance (accuracy scores, in percentages) in selected CLiMP (top part) and SLING (bottom part) paradigms. In each row (paradigm), four highest scores are highlighted in bold.

3 Results and analyses

3.1 Model performance

For the majority of paradigms in both benchmarks, we do not observe consistent differences between monolingual and bilingual models' scores (see Tables A1–A2 in the Appendix). This result is expected, due to large variation in model architectures and the amounts of data they are trained on.

At the same time, from Table 2 we see that 3 out of 16 CLiMP paradigms and 4 out of 39 SLING paradigms yield very consistent differences between bilingual and monolingual model scores, and for 9 more SLING paradigms the differences are consistent except the inexplicably low performance of the monolingual PANGU model. Adding up these numbers, we observe reliable differences in 16 out of the 55 paradigms (29%). To compute how likely this result may occur by chance, we use bootstrapping, randomly sampling two sets of four scores (in the range between 0.00 and 100.00) 55 times to see whether we obtain the result like ours or more extreme (i.e., for at least 7 cases all scores in one set are greater than those in the other set, and for at least 9 more cases three scores from one set are greater than all scores in the other set). Having repeated this process 100k times, we estimate the probability of obtaining a result like ours (or more extreme) to be 0.069%, a very low value.

Out of the 16 paradigms with consistent differences, bilingual models show higher scores in 4 paradigms, indicating either positive cross-lingual transfer or the bilingual models' advantage because of their larger sizes. The monolingual models are better in 12 paradigms, indicating negative transfer. In other words, these results suggest that negative cross-lingual transfer is common in bilingual language models, and having English training data in addition to Chinese is not always helpful.

We have shown that monolingual models (including the much smaller BERT) score better than bilingual models on some linguistic paradigms, and that this result is statistically unlikely. We now turn to analyzing the profiles of models' feature attribution to answer the question: Can the different scores of monolingual vs. bilingual models be explained by the differences in how well they understand the key relations between words in target sentences?

3.2 Feature attribution analysis

We investigate how the important words from the left context affect the generation of the target word in the sentences from the two evaluation benchmarks above. Consider again example (1) from Section 2.2. After reading the last word 制片人 ('producer'), a human speaker should note the presence of the word 还是 ('or'), which indicates an alternative question and calls for the end of sentence rather than the 吗 (*ma*) particle. Analogously, in the context of LLMs, after decoding 制片人 ('producer'), to generate an appropriate token, the model should focus on the token 还是 ('or'). This *keyword* suggests that the end of sentence (in this case, a question mark) is a more appropriate token to generate than the 吗 (*ma*) particle. Consequently, we expect a (monolingual) model with higher performance on the paradigm represented by this sentence to assign a higher importance value to the keyword (here: 还是, 'or') during the generation of a target token (here: question mark), compared to a (bilingual) model with lower performance.

To test this hypothesis, we use the Inseq interpretability toolkit (Sarti et al., 2023), which is well suited to analyze feature attribution using gradient-based methods. Given the left context, we force a model to decode the next target token from the grammatical sentence (the question mark in the example above). We then use the integrated gradients method to compute the distribution of importance scores for all preceding tokens and extract the (normalized) score for the keyword (还是, 'or', in the example above). Finally, we compare the scores for a monolingual and a bilingual model.

We focus on one monolingual (CPM) and one bilingual model (YI), thanks to their Inseq support. Furthermore, we only consider two SLING paradigms (*Anaphor gender: baseline female* and *Alternative question: haishi ma*), as the rest were either incompatible with left-to-right processing (i.e., generating the correct target token requires right sentence context) or yielded tokenization patterns of the keyword and/or the target token that were different across the two models (CPM and YI), which would generate multiple scores per word and

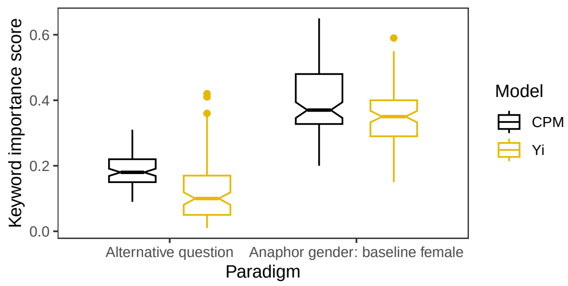

Figure 1: Keyword importance scores of the monolingual CPM and bilingual YI model in two paradigms.

possibly render the comparison unfair. For each paradigm, we consider the first 100 grammatical sentences. We extract the importance scores as described above for both models, compare their means and test whether there are statistically significant differences using the Wilcoxon signed-rank test (Wilcoxon, 1992) while correcting for false discovery rate (Benjamini and Hochberg, 1995).

From Figure 1, we see that in both paradigms the monolingual model yields higher keyword importance scores than the bilingual one. Our statistical tests confirm that the differences are significant, with both $p < .001$. This suggests that, indeed, the monolingual CPM model better captures the relations between the keyword and the target token, which can explain its higher performance on some paradigms compared to the bilingual YI model.

4 Conclusion

We have evaluated four monolingual Chinese and four bilingual Chinese–English models on two Chinese linguistic benchmarks. Monolingual models, despite their smaller sizes, consistently outperform bilingual ones on a number of paradigms, a statistically unlikely result which suggests that bilingual Chinese–English models suffer from negative cross-lingual transfer. This extends the results on negative transfer generally reported for *multilingual* models (Chang et al., 2023) to a bilingual setting where a model is trained on two high-resource languages well-represented in models' training data. Our findings have implications for the ongoing effort of training bilingual LLMs on high-resource languages (Faysse et al., 2024; Zhang et al., 2024; Nikolich et al., 2024, etc.). Our feature attribution analysis suggests that the higher scores of the monolingual models may have to do with their better (compared to the bilingual models) understanding of the key relations between words.

5 Limitations

This study only focuses on one language pair, English and Chinese, and only one direction of cross-lingual transfer (English to Chinese). It is unclear whether the results would generalize to other language pairs or to transfer from Chinese to English. We only consider a total of eight LLMs, all with 14B parameters or less, and the results may be different for larger models. The models we have compared differ on many dimensions, including architecture, size, objective, while ideally one would compare a monolingual and a bilingual model that only differ in their training data (one vs. two languages), to focus on the impact of bilingual training. The benchmarks we use, CLiMP and SLING, also come with limitations, namely they only evaluate the models' linguistic knowledge. Our interpretability analysis is further limited to only two paradigms, a constraint imposed by our method's requirement of left-to-right processing and by different tokenization schemes used in the models.

As we only evaluate existing models, we do not anticipate any risks related to misuse or negative application of the results presented in our study. However, our focus on the two languages with the highest amount of training data available contributes to underexposure of lower-resource languages.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Baichuan. 2023. Baichuan-7b.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *Preprint*, arXiv:2401.10440.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2023. When is multilinguality a curse? Language modeling for 250 high-and low-resource languages. *Preprint*, arXiv:2311.09205.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, Antonio Loison, Duarte Alves, Caio Corro, Nicolas Boizard, Jaoe Alves, Ricardo Rei, Pedro Raphaël Martins, Antoni Casademunt, François Yvon, André Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. CroissantLLM: A Truly Bilingual French-English Language Model. *Preprint*, arXiv:2402.00786.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for Russian. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 189–199.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *Preprint*, arXiv:2107.02137.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *Preprint*, arXiv:2309.10305.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 All Tools. *Preprint*, arXiv:2406.12793.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. PanGu-α: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation. *Preprint*, arXiv:2104.12369.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024. Map-neo: Highly capable and transparent bilingual large language model series. *Preprint*, arXiv:2405.19327.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. 2021. Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, 2:93–99.

A Appendix A. Detailed evaluation scores

Paradigm	Monolingual models				Bilingual models			
	ERNIE	CPM	PANGU	BERT	QWEN	BAICHUAN	YI	CHATGLM
Anaphor agreement								
—”— gender	**85.6**	79.9	**92.6**	**86.2**	64.0	**86.5**	62.5	77.4
Binding								
—”— gender	**54.2**	51.3	**61.2**	50.8	50.0	**58.6**	51.2	**81.0**
ba construction								
—”—	**63.0**	57.8	19.3	**69.0**	62.4	**74.3**	**73.5**	60.7
Coverb								
—”— instrument	57.5	36.0	54.1	**91.1**	**80.8**	**79.5**	**80.5**	79.0
—”— with	82.3	61.7	73.5	84.7	**86.2**	**84.9**	**84.8**	**84.8**
NP head finality								
—”— clause	67.1	**86.5**	65.6	53.1	**80.3**	76.8	**80.6**	**80.2**
Classifier								
—”—	85.8	57.1	76.0	**95.6**	**92.4**	**90.2**	**90.2**	**93.8**
—”— adj	**87.8**	55.5	69.1	**93.2**	**91.8**	84.2	87.0	**88.1**
—”— clause	**84.3**	52.2	66.5	**90.0**	**89.3**	80.8	**84.3**	80.9
Filler gap								
—”— dependency	**87.3**	62.3	**91.9**	62.4	**71.1**	65.2	**70.3**	64.9
Passive								
—”— formal	**60.9**	47.0	**61.6**	**67.1**	53.8	50.3	49.2	**60.2**
Verb complement								
—”— direction	**96.2**	81.4	80.1	**93.0**	85.0	**91.8**	**86.1**	84.0
—”— duration	**92.8**	83.6	82.6	**90.2**	89.7	**92.8**	**94.2**	86.9
—”— frequency	**98.4**	48.8	**75.6**	**97.8**	19.9	25.4	32.6	**81.3**
—”— res adj	59.7	25.9	59.3	87.6	**92.1**	**95.2**	**91.1**	**90.9**
—”— res verb	**92.8**	**96.7**	**90.1**	**96.2**	61.2	65.7	61.4	64.2

Table A1: The models’ performance (accuracy scores, in percentages) on CLiMP paradigms. Four highest scores in each paradigm are highlighted in boldface.

Paradigm	Monolingual models				Bilingual models			
	ERNIE	CPM	PANGU	BERT	QWEN	BAICHUAN	YI	CHATGLM
Alternative Question								
haishi ma	**94.6**	**85.8**	10.0	**93.1**	9.8	26.6	6.5	**64.0**
Anaphor (Gender)								
baseline female	**92.9**	**89.8**	**95.9**	**86.7**	32.1	66.2	70.3	67.1
baseline male	30.4	**53.8**	**100.0**	46.1	**48.9**	34.7	47.7	**64.5**
pp female	59.1	**95.2**	**98.6**	**87.0**	77.3	**96.3**	69.6	78.3
pp male	38.8	46.3	**99.9**	**76.0**	**79.8**	21.0	73.8	**74.2**
self female	92.8	66.4	**97.3**	93.3	**100.0**	**99.4**	**97.2**	90.4
self male	70.7	**86.7**	**100.0**	**88.4**	0.1	**75.0**	21.0	47.4
Anaphor (Number)								
baseline cl female	**99.5**	**77.9**	0.0	**99.4**	10.1	16.2	29.4	**40.7**
baseline cl male	**99.9**	**75.1**	0.0	**99.6**	26.0	42.9	**47.6**	45.3
baseline cl men female	**99.5**	**88.8**	0.0	**99.4**	5.9	9.7	25.3	**34.8**
baseline cl men male	**100.0**	**87.6**	0.0	**100.0**	17.9	38.0	38.9	**43.2**
baseline men female	**99.3**	**51.8**	0.0	**98.0**	6.7	9.4	28.7	**41.4**
baseline men male	**99.7**	**49.5**	0.1	**99.7**	20.2	40.4	41.1	**52.8**
cl men self female	**98.3**	**96.2**	0.0	**100.0**	87.5	**95.4**	84.0	77.9
cl men self male	**99.6**	97.1	0.0	**100.0**	**100.0**	**99.7**	98.8	93.3
cl self female	**99.2**	**88.8**	0.0	**99.9**	74.8	**82.8**	62.4	70.2
cl self male	**99.5**	85.8	0.1	**99.9**	**100.0**	96.3	**97.5**	92.2
manself female	**96.1**	67.4	0.0	**98.8**	**89.2**	**83.4**	80.5	61.3
manself male	98.3	61.1	0.0	**99.3**	**100.0**	**98.7**	**98.7**	94.3
Aspect								
temporal guo	**91.8**	79.7	72.4	**95.5**	81.3	82.8	**92.1**	**93.2**
temporal le	59.7	**78.8**	**73.9**	65.2	63.2	64.8	**70.5**	**74.6**
zai guo	**92.0**	78.6	65.4	**97.9**	77.5	**87.6**	**79.7**	79.4
zai no le	**64.8**	0.8	16.1	**85.2**	53.8	50.0	**57.0**	**59.4**
Classifier-Noun								
cl adj comp noun	**69.7**	55.6	53.4	70.7	**66.4**	**66.1**	**64.4**	63.0
cl adj comp noun v2	**85.5**	46.0	50.7	**87.5**	70.6	**71.9**	**76.8**	62.8
cl adj simple noun	**93.1**	58.9	77.1	**96.5**	92.8	**92.9**	**93.0**	79.8
cl comp noun	65.6	51.0	53.8	**69.8**	**62.9**	**68.8**	59.7	**67.6**
cl comp noun v2	**85.1**	45.2	55.5	**86.7**	70.2	70.0	**78.2**	**76.8**
cl simple noun	**96.1**	61.2	85.0	**98.5**	**96.0**	**95.1**	94.7	88.4
dem cl swap	**99.5**	52.5	85.7	**99.8**	88.7	**92.1**	**92.7**	88.7
Definiteness Effect								
demonstrative	**93.9**	48.3	49.3	**98.2**	**83.4**	58.0	44.5	**70.6**
every	**96.2**	**92.5**	87.7	**94.6**	**88.0**	69.2	58.7	84.9
Polarity Item								
any	85.2	**95.9**	**93.6**	65.8	82.9	**92.1**	77.2	**95.4**
even wh	85.8	42.3	47.7	52.4	**97.7**	**98.4**	**96.9**	**98.0**
more or less	**98.3**	**98.6**	**97.6**	**97.9**	86.2	96.8	93.3	79.5
Relative Clause								
rc resumptive noun	15.2	**82.1**	16.7	25.6	**37.9**	**25.8**	**31.4**	24.7
rc resumptive pronoun	54.8	18.6	11.8	42.7	**64.3**	**77.8**	**68.1**	**60.8**
Wh-fronting								
bare wh	**100.0**	96.6	99.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
mod wh	**100.0**	90.7	88.8	99.5	**100.0**	**100.0**	**99.9**	99.6

Table A2: The models' performance (accuracy scores, in percentages) on SLING paradigms. Four highest scores in each paradigm are highlighted in boldface.