Electronic Supplementary Material

Table S1 Performance of CELLM in C-Eval (Huang et al., 2024).

Case	Metric	Accuracy value
C-Eval-computer_network	Accuracy	36.84
C-Eval-operating_system	Accuracy	31.58
C-Eval-computer_architecture	Accuracy	19.05
C-Eval-college_programming	Accuracy	32.43
C-Eval-college_physics	Accuracy	5.26
C-Eval-college_chemistry	Accuracy	20.83
C-Eval-advanced_mathematics	Accuracy	21.05
C-Eval-probability_and_statistics	Accuracy	22.22
C-Eval-discrete_mathematics	Accuracy	18.75
C-Eval-electrical_engineer	Accuracy	27.03
C-Eval-metrology_engineer	Accuracy	20.83
C-Eval-high_school_mathematics	Accuracy	22.22
C-Eval-high_school_physics	Accuracy	26.32
C-Eval-high_school_chemistry	Accuracy	10.53
C-Eval-high_school_biology	Accuracy	21.05
C-Eval-middle_school_mathematics	Accuracy	26.32
C-Eval-middle_school_biology	Accuracy	19.05
C-Eval-middle_school_physics	Accuracy	15.79
C-Eval-middle_school_chemistry	Accuracy	15.00
C-Eval-veterinary_medicine	Accuracy	17.39
C-Eval-college_economics	Accuracy	25.45
C-Eval-business_administration	Accuracy	24.24
C-Eval-marxism	Accuracy	26.32
C-Eval-mao_zedong_thought	Accuracy	20.83
C-Eval-education_science	Accuracy	34.48

C-Eval-teacher_qualification	Accuracy	15.91
C-Eval-high_school_politics	Accuracy	26.32
C-Eval-high_school_geography	Accuracy	31.58
C-Eval-middle_school_politics	Accuracy	33.33
C-Eval-middle_school_geography	Accuracy	25.00
C-Eval-modern_chinese_history	Accuracy	26.09
C-Eval-ideological_and_moral_cultivation	Accuracy	26.32
C-Eval-logic	Accuracy	27.27
C-Eval-law	Accuracy	33.33
C-Eval-chinese_language_and_literature	Accuracy	26.09
C-Eval-art_studies	Accuracy	18.18
C-Eval-professional_tour_guide	Accuracy	20.69
C-Eval-legal_professional	Accuracy	43.48
C-Eval-high_school_chinese	Accuracy	15.79
C-Eval-high_school_history	Accuracy	30.00
C-Eval-middle_school_history	Accuracy	27.27
C-Eval-civil_servant	Accuracy	23.40
C-Eval-sports_science	Accuracy	31.58
C-Eval-plant_protection	Accuracy	18.18
C-Eval-basic_medicine	Accuracy	31.58
C-Eval-clinical_medicine	Accuracy	22.73
C-Eval-urban_and_rural_planner	Accuracy	28.26
C-Eval-accountant	Accuracy	32.65
C-Eval-fire_engineer	Accuracy	16.13
C-Eval-environmental_impact_assessment_engineer	Accuracy	22.58
C-Eval-tax_accountant	Accuracy	14.29
C-Eval-physician	Accuracy	16.33
C-Eval-stem	Accuracy	21.48
C-Eval-social-science	Accuracy	26.35

C-Eval-humanities	Accuracy	26.77
C-Eval-other	Accuracy	23.43
C-Eval-hard	Accuracy	18.40
C-Eval	Accuracy	23.95

Note. CELLM: excel extension large language model.

Table S2 Perfomance of CELLM in CMMLU (Li et al., 2023)

Case	Metirc	Accuracy value
CMMLU-agronomy	Accuracy	25.44
CMMLU-anatomy	Accuracy	23.65
CMMLU-ancient-chinese	Accuracy	27.44
CMMLU-arts	Accuracy	31.25
CMMLU-astronomy	Accuracy	23.03
CMMLU-business_ethics	Accuracy	22.01
CMMLU-chinese_civil_service_exam	Accuracy	30.63
CMMLU-chinese_driving_rule	Accuracy	31.30
CMMLU-chinese_food_culture	Accuracy	27.94
CMMLU-chinese_foreign_policy	Accuracy	26.17
CMMLU-chinese_history	Accuracy	24.77
CMMLU-chinese_literature	Accuracy	25.00
CMMLU-chinese_teacher_qualification	Accuracy	22.91
CMMLU-clinical_knowledge	Accuracy	23.21
CMMLU-college_actuarial_science	Accuracy	14.15
CMMLU-college_education	Accuracy	19.63
CMMLU-college_engineering_hydrology	Accuracy	19.81
CMMLU-college_law	Accuracy	21.30

CMMLU-college_mathematics	Accuracy	22.86
CMMLU-college_medical_statistics	Accuracy	24.53
CMMLU-college_medicine	Accuracy	24.91
CMMLU-computer_science	Accuracy	21.08
CMMLU-computer_security	Accuracy	26.90
CMMLU-conceptual_physics	Accuracy	30.61
CMMLU-construction_project_management	Accuracy	19.42
CMMLU-ecnomics	Accuracy	23.27
CMMLU-education	Accuracy	19.02
CMMLU-electrical_engineering	Accuracy	28.49
CMMLU-elementary_chinese	Accuracy	29.76
CMMLU-elementary_commonsense	Accuracy	26.77
CMMLU-elementary_information_and_technology	Accuracy	23.53
CMMLU-elementary_mathematics	Accuracy	25.22
CMMLU-ethnology	Accuracy	22.22
CMMLU-food_science	Accuracy	25.17
CMMLU-genetics	Accuracy	25.00
CMMLU-global_facts	Accuracy	22.15
CMMLU-high_school_biology	Accuracy	20.12
CMMLU- high_school_chemistry	Accuracy	23.48
CMMLU- high_school_geography	Accuracy	22.03
CMMLU- high_school_mathematics	Accuracy	19.51
CMMLU- high_school_physics	Accuracy	17.27
CMMLU- high_school_politics	Accuracy	28.67
CMMLU-human_sexuality	Accuracy	26.19
CMMLU-international_law	Accuracy	29.19
CMMLU-journalism	Accuracy	26.16
CMMLU-jurisprudence	Accuracy	25.30
CMMLU-legal_and_moral_basis	Accuracy	25.70

CMMLU-logical	Accuracy	20.33
CMMLU-machine_learning	Accuracy	22.13
CMMLU-management	Accuracy	26.67
CMMLU-marketing	Accuracy	25.56
CMMLU-marxist_theory	Accuracy	25.93
CMMLU-modern_chinese	Accuracy	25.00
CMMLU-nutrition	Accuracy	28.28
CMMLU-philosophy	Accuracy	26.67
CMMLU-professional_accounting	Accuracy	24.00
CMMLU-professional_law	Accuracy	28.44
CMMLU-professional_medicine	Accuracy	20.48
CMMLU-professional_psychology	Accuracy	27.16
CMMLU-public_relations	Accuracy	18.97
CMMLU-security_study	Accuracy	23.70
CMMLU-sociology	Accuracy	29.20
CMMLU-sports_science	Accuracy	22.42
CMMLU-traditional_chinese_medicine	Accuracy	24.86
CMMLU-virology	Accuracy	22.49
CMMLU-world_history	Accuracy	29.19
CMMLU-world_religions	Accuracy	31.87
CMMLU-humanities	Average	26.26
CMMLU-stem	Average	22.55
CMMLU-social-science	Average	24.91
CMMLU-other	Average	24.97
CMMLU-china-specific	Average	26.19
CMMLU	Average	24.59

Notes. CELLM: excel extension large language model, CMMLU: measuring massive multitask language understanding in Chinese.

Table S3 Performance of CELLM in MMLU (Hendrycks et al., 2021a)

Case	Metirc	Accuracy value
MMLU_college_biology	Accuracy	25.44
MMLU_college_chemistry	Accuracy	23.65
MMLU_college_computer_science	Accuracy	27.44
MMLU_college_mathematics	Accuracy	31.25
MMLU_college_physics	Accuracy	23.03
MMLU_electrical_engineering	Accuracy	22.01
MMLU_astronomy	Accuracy	30.63
MMLU_anatomy	Accuracy	31.30
MMLU_abstract_algebra	Accuracy	27.94
MMLU_machine_learning	Accuracy	26.17
MMLU_clinical_knowledge	Accuracy	24.77
MMLU_global_facts	Accuracy	25.00
MMLU_management	Accuracy	22.91
MMLU_nutrition	Accuracy	23.21
MMLU_marketing	Accuracy	14.15
MMLU_professional_accounting	Accuracy	19.63
MMLU_high_school_geography	Accuracy	19.81
MMLU_international_law	Accuracy	21.30
MMLU_moral_scenarios	Accuracy	22.86
MMLU_computer_security	Accuracy	24.53
MMLU_high_school_microeconomics	Accuracy	24.91
MMLU_professional_law	Accuracy	21.08
MMLU_medical_genetics	Accuracy	26.90
MMLU_professional_psychology	Accuracy	30.61

MMLU_jurisprudence	Accuracy	19.42
MMLU_world_religions	Accuracy	23.27
MMLU_philosophy	Accuracy	19.02
MMLU_virology	Accuracy	28.49
MMLU_high_school_chemistry	Accuracy	29.76
MMLU_public_relations	Accuracy	26.77
MMLU_high_school_macroeconomics	Accuracy	23.53
MMLU_huamn_sexuality	Accuracy	25.22
MMLU_elementary_mathematics	Accuracy	22.22
MMLU_high_school_physics	Accuracy	25.17
MMLU_high_school_computer_science	Accuracy	25.00
MMLU_high_school_european_history	Accuracy	22.15
MMLU_business_ethics	Accuracy	20.12

Notes. CELLM: excel extension large language model, MMLU: measuring massive multitask language understanding.

Table S4 Performance of CELLM in traditional natural language processing tasks, including CLUE and Lambada (Paperno et al., 2016; Xu et al., 2020).

Case	Metric	Accuracy value
OCNLI	Accuracy	30.20
AFQMC-dev	Accuracy	29.12
C3	Accuracy	15.89
CMNLI	Accuracy	32.60
DRCD_dev	Accuracy	0.28
Lambada	Accuracy	2.72

Notes. CELLM: excel extension large language mode, CLUE: Chinese language understanding evaluation benchmark, OCNLI :original Chinese natural language inference, AFQMC: auxiliary-field quantum monte carlo, CMNLI: Chinese multi-genre natural language inference, DRCD: delta reading comprehension dataset.

Table S5 Performance of CELLM in wikibench (Kuo et al., 2024)

Metric	Accuracy value
acc_4	20.75
acc_1	19.40
more_1_0	100.00
more_1_1	19.40
more_4_0	100.00
more_4_1	72.85
more_4_2	9.65
more_4_3	0.45
more_4_4	0.05
perf_1	19.40
perf_4	0.05
vote_4	19.05
vote_1	19.40
prior_A	51.98
prior_B	15.64
prior_C	6.05
prior_D	9.50
prior_-	16.84

Note. CELLM: excel extension large language model.

Table S6 Performance of CELLM in Arc (Clark et al., 2018)

Case name	Metric	Accuracy value
Arc-easy	Accuracy	21.69
Arc-challenge	Accuracy	20.68

Note. CELLM: excel extension large language model.

Table S7 Performance of CELLM in LCSTS (Hu et al., 2015)

Metric	Accuracy value
Rouge 1	8.71
Rouge 2	2.00
Rouge L	6.77

Note. CELLM: excel extension large language model, LCSTS: large scale Chinese short text summarization dataset.

Table S8 Performance of CELLM in Gaokao-bench (Zhang et al., 2023)

Case	Metric	Accuracy value
GaokaoBench_2010-2022_Math_I_MCQs	Score	9.38
GaokaoBench_2010-2022_History_I_MCQs	Score	26.39
GaokaoBench_2010-2022_Biology_I_MCQs	Score	15.62
GaokaoBench_2010-2022_Political_Science_MCQs	Score	30.00
GaokaoBench_2010-2022_Physics_MCQs	Score	2.34
GaokaoBench_2010-2022_English_MCQs	Score	20.00
GaokaoBench_2010-2022_Chinese_Modern_Lit	Score	12.50
GaokaoBench_2010-2022_English_Fill_in_Blanks	Score	9.50
GaokaoBench_2012-2022_English_Cloze_Test	Score	5.38
GaokaoBench_2010-2022_Geography_MCQs	Score	9.30
GaokaoBench_2010-2022_English_Reading_Comp	Score	7.87

Notes. CELLM: excel extension large language model, MCQs: multiple choice

questions, Comp: competition.

Table S9 Performance of CELLM in AGI-Eval (Zhong et al., 2024)

Case	Metirc	Accuracy value
AGI-Eval-gaokao-chinese	Accuracy	21.14
AGI-Eval-gaokao-english	Accuracy	20.59
AGI-Eval-gaokao-geography	Accuracy	8.54
AGI-Eval-gaokao-history	Accuracy	18.30
AGI-Eval-gaokao-biology	Accuracy	21.43
AGI-Eval-gaokao-chemistry	Accuracy	19.32
AGI-Eval-gaokao-mathqa	Accuracy	10.54
AGI-Eval-logiqa-zh	Accuracy	17.67
AGI-Eval-lsat-ar	Accuracy	20.43
AGI-Eval-lsat-lr	Accuracy	19.80
AGI-Eval-lsat-rc	Accuracy	18.96
AGI-Eval-logiqa-en	Accuracy	24.27
AGI-Eval-sat-math	Accuracy	18.64
AGI-Eval-sat-en	Accuracy	16.50
AGI-Eval-sat-en-without-passage	Accuracy	21.36
AGI-Eval-aqua-rat	Accuracy	16.89
AGI-Eval-gaokao-physics	Accuracy	12.00
AGI-Eval-jec-qa-ca	Accuracy	10.70
AGI-Eval-gaokao-mathcloze	Score	0.00
AGI-Eval-math	Score	0.90
AGI-Eval-chinese	Accuracy	14.19
AGI-Eval-english	Accuracy	17.53

AGI-Eval-gaokao	Accuracy	14.65
AGI-Eval	Accuracy	15.62

Notes. CELLM: excel extension large language model, AGI: artificial general intelligence.

References

Hu, B. T., Chen, Q. C., & Zhu, F. Z. (2015). LCSTS: A large scale Chinese short text summarization dataset. In: *Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon: Association for Computational Linguistics, 1967−1972.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., & Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics, 1525–1534.

Zhang, X. T., Li, C. Y., Zong, Y., Ying, Z. Y., He, L., & Qiu, X. P. (2023). Evaluating the performance of large language models on GAOKAO benchmark. *arXiv Preprint*, arXiv:2305.12474.