Appendix

A APPENDIX DESCRIPTION

In Sec. B and Sec. C, we elaborate on the specific models employed in our study. Additionally, Sec. D provides a comprehensive evaluation of these models' performance across various tasks, including detailed scoring metrics. Moreover, the prompt templates utilized for these models are also presented in Sec. E.

B LLM EMPLOYED

Below is the LLMs we utilized in our experiment:

- DeepSeek-R1-8B [5]: Developed by DeepSeek, this model has 8 billion parameters and is designed for general-purpose language understanding and generation tasks. It is built on top of the Llama architecture and has been fine-tuned for improved performance in a variety of natural language processing applications.
- DeepSeek-R1-7B [5]: A smaller version of the DeepSeek-R1 family with 7 billion parameters, it offers a balance between model size and performance, making it suitable for applications where computational resources are limited.
- DeepSeek-R1-1.5B [5]: The smallest variant in the DeepSeek-R1 series, with 1.5 billion parameters. It provides a more lightweight option for language processing tasks while maintaining reasonable performance.
- Qwen2.5-7B [8]: Created by Moonshot AI, Qwen2.5-7B is a 7-billion parameter model that focuses on offering high-quality text generation and comprehension capabilities. It is designed to handle various language-related tasks efficiently.
- Llama2-8B [6]: Developed by Meta, Llama 2 is a family of large language models. The 8B variant has 8 billion parameters and is designed to be more accessible to researchers and developers. It has been trained to be more aligned with human values and has improved safety features.
- ChatGLM4-9B [11]: Developed by the GLM team, this model has 9 billion parameters and is specifically designed for dialogue applications. It is built on the GLM (General Language Model) architecture and has been optimized for interactive conversations.
- Gemma9B [4]: Developed by Google, Gemma9B have 9 billion parameters. It focuses on providing comprehensive language understanding and generation capabilities.
- AIDC7B [1]: Developed by the AI Research Center of DAMO Academy, this model has 7 billion parameters and is designed to support a wide range of natural language processing tasks, including text generation, comprehension, and translation.
- Seed7B [3]: Created by the AI team of ByteDance, Seed7B is a 7-billion parameter model that aims to provide efficient and effective language processing solutions for various applications.
- InternLM2.5-7B [7]: Developed by the AI team of Shanghai AI Laboratory, this model has 7 billion parameters and is designed to offer strong language understanding and generation capabilities. It has been trained on a diverse dataset to ensure robust performance across different domains.
- ERNIE3.5 [10]: Baidu's ERNIE 3.5 is a pre-trained language model that has been fine-tuned on various tasks. It has been designed to better understand the semantic relationships in text and has shown strong performance in multiple natural language processing benchmarks.
- Qwen2.5-Coder-7B [8]: A specialized variant of the Qwen2.5 family, this 7-billion parameter model is tailored for coding-related tasks. It has been trained on a large corpus of code to assist with code generation, comprehension, and debugging.
- Qwen2.5-1.5B [8]: The smallest model in the Qwen2.5 series with 1.5 billion parameters. It provides a more lightweight option for general language tasks while maintaining the quality of text generation.
- Yi1.5-9B [2]: Developed by the AI team of 01.AI, Yi1.5-9B is a 9-billion parameter model that focuses on delivering high-quality language generation and understanding. It is designed to be more efficient and accessible for a variety of applications.
- Qwen2-7B [8]: Another variant in the Qwen family, this 7-billion parameter model continues to build on the capabilities of its predecessors, offering improved performance in language generation and comprehension tasks.
- Yi1.5-6B [2]: A smaller variant of the Yi family with 6 billion parameters, it provides a balance between model size and performance, making it suitable for applications with moderate computational resources.

C VLM EMPLOYED

Below is the VLMs we utilized in our experiment:

- Doubao-vision1.5-Pro [3] Doubao-vision1.5-Pro is a VLM excels in visual reasoning, document recognition, and fine-grained information understanding. It also features a highly efficient inference system, leveraging heterogeneous hardware and low-precision optimization strategies to achieve low latency and high throughput.
- BaichuanGLM4VPlus [11] BaichuanGLM4VPlus is a multimodal language model that integrates vision and language capabilities. It is designed to provide accurate and context-aware responses to queries involving both text and images. The model is optimized for tasks such as image description, visual question answering, and document understanding.
- Qwen2VL7B2 [8] Qwen2VL7B2 is a 7.2 billion parameter multimodal model that supports high-resolution image understanding and video processing. It features dynamic resolution handling and multimodal rotary position embedding (M-ROPE) to capture spatial and temporal information effectively. The model is designed for applications such as visual question answering, video content creation, and agent-based tasks.

- BaichuanGLM4VFlash [11] BaichuanGLM4VFlash is a lightweight multimodal model optimized for fast inference and deployment. It provides efficient processing of visual and textual data, making it suitable for real-time applications and mobile devices. The model is designed to deliver accurate results with minimal computational resources.
- Qwen2VL72B [8] Qwen2VL72B is a large-scale multimodal model with 72 billion parameters. It demonstrates state-of-the-art performance in visual understanding, video processing, and multilingual support. The model is capable of handling complex visual tasks and generating high-quality responses for a wide range of applications.
- Qwen2VL7B [8] Qwen2VL7B is a 7 billion parameter multimodal model that offers a balance between performance and efficiency. It supports various visual tasks, including document understanding, video question answering, and multilingual processing. The model is designed to provide accurate and context-aware responses to multimodal queries.
- Doubao-vision-pro-32k [3] Doubao-vision-pro-32k is a multimodal model with a 32k context window, supporting long text and high-resolution image inputs. It excels in visual reasoning, document recognition, and fine-grained information extraction. The model is designed to handle complex visual tasks and provide accurate responses in various applications.
- Doubao-vision-lite [3] Doubao-vision-lite is a lightweight version of the Doubao-vision model, optimized for efficient inference and deployment. It maintains high performance in visual tasks while reducing computational requirements, making it suitable for resource-constrained environments.
- DeepSeekVL2 [9] DeepSeekVL2 is a series of advanced MoE visual-language models that offer significant improvements over their predecessors. The models support dynamic tiling for high-resolution images, multilingual OCR, and efficient processing of complex visual data. They are designed for tasks such as visual question answering, document understanding, and visual localization.

D DETAILED INFORMATION ON RECOGNIZED MODELS

Below we list the evaluated models in each task using their official full names.

E PROMPT TEMPLATES FOR LARGE MODELS

Below are the exact prompts used. Each elicits a single-character response (“T” or “F”) with no extra text.

E.1 Number Comparison

```
Compare NUM1:{i} vs NUM2:{j}. Strict rules:
  1. Treat both as integers.
  2. Apply standard numerical comparison.
  3. If NUM1 < NUM2, output T; if NUM1 > NUM2, output F.
  4. The first letter must be T or F.
Do not output any explanation or additional text.
```

Table 1: Accuracy on Historical Reasoning Task

Model Name	Accuracy
DeepSeek-R1-8B	0.77570
DeepSeek-R1-7B	0.67256
DeepSeek-R1-1.5B	0.65094
Qwen2.5-7B	0.60714
Llama2-8B	0.57894
ChatGLM4-9B	0.57692
Gemma9B	0.57407
AIDC7B	0.56140
Seed7B	0.54716
InternLM2.5-7B	0.54385
ERNIE3.5	0.53448
Qwen2.5-Coder-7B	0.52777
Qwen2.5-1.5B	0.50925
Yi1.5-9B	0.49514
Qwen2-7B	0.45714
Yi1.5-6B	0.41666

Table 2: Accuracy on Numerical Comparison Task

Model Name	Accuracy
AIDC7B	0.67326
internLM2.5-7B	0.65656
DeepSeek-R1-7B	0.62589
Gemma9B	0.61313
ERNIE3.5	0.48900
Qwen2.5-1.5B	0.48837
DeepSeek-R1-8B	0.47407
ChatGLM4-9B	0.43884
Qwen2.5-Coder-7B	0.41843
Llama2-8B	0.41353
DeepSeek-R1-1.5B	0.39166
Yi1.5-6B	0.39716
Yi1.5-9B	0.39716
Qwen2-7B	0.39716
Qwen2.5-7B	0.36666
Seed7B	0.36170

Table 3: Accuracy on Letter-Frequency Comparison Task

Model Name	Accuracy
DeepSeek-R1-1.5B	0.67619
Qwen2-7B	0.65693
Qwen2.5-7B	0.63917
AIDC7B	0.62886
DeepSeek-R1-8B	0.61666
Llama2-8B	0.61538
internLM2.5-7B	0.61052
Qwen2.5-1.5B	0.60784
DeepSeek-R1-7B	0.60759
Seed7B	0.59677
Qwen2.5-Coder-7B	0.52884
ChatGLM4-9B	0.40659

Table 4: Accuracy on Visual Age Comparison Task

Model Name	Accuracy
Doubao-vision1.5-Pro	0.70058
BaichuanGLM4VPlus	0.69528
Qwen2VL7B2	0.66935
BaichuanGLM4VFlash	0.59915
Qwen2VL72B	0.43644
Qwen2VL7B	0.43037
Doubao-vision-pro-32k	0.31034
Doubao-vision-lite	0.38000
DeepSeekVL2	0.29411

E.2 Historical Chronology

```
Compare the chronological order of two events:
  Event1: {i}, Event2: {j}. Strict rules:
  1. Use historical facts.
  2. Output T if Event1 occurred first; F otherwise.
  3. The first letter must be T or F.
Do not output any explanation or additional text.
```

E.3 Letter-Frequency Comparison

```
Compare TEXT1:{i} vs TEXT2:{j}. Strict rules:
  1. Count occurrences of 'r' only.
  2. If TEXT1 has more 'r', output F;
     if TEXT2 has more 'r', output T.
  3. Respond with a single 'T' or 'F'.
  4. No explanation or extra text.
```

E.4 Visual Age Comparison

```
Analyze facial features, skin texture and age
of both subjects.
Compare apparent ages:
  - If left appears older, output T
  - If right appears older, output F
Respond only with single character [T/F]
Do not include explanations or extra text.
```

REFERENCES

[1] [n. d.]. AIDC7B. https://huggingface.co/AIDC-AI/Marco-o1

[2] 01.AI. 2024. Yi: Open Foundation Models by 01.AI. arXiv preprint arXiv:2403.04652 (2024). https://arxiv.org/abs/2403.04652

[3] ByteDanceAITeam. 2024. ByteDanceDoubaoLLMVLM. https://team.doubao.com/en/special/doubao_1_5_pro

[4] Google DeepMind. 2024. Gemma: Open Models Based on Gemini Research and Technology. https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf

[5] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[6] OpenAI. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288 (2023). https://arxiv.org/abs/2307.09288

[7] InternLM Team. 2024. InternLM2 Technical Report. arXiv preprint arXiv:2403.17297 (2024). https://arxiv.org/abs/2403.17297

[8] Qwen Team. [n. d.]. Qwen2.5 Technical Report. Qwen. https://arxiv.org/abs/2412.15115

[9] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, Z. Xie, Y. Wu, K. Hu, J. Wang, Y. Sun, Y. Li, Y. Piao, K. Guan, A. Liu, X. Xie, Y. You, K. Dong, X. Yu, H. Zhang, L. Zhao, Y. Wang, and C. Ruan. 2024. DeepSeekVL2: Mixture-of-experts vision-language models for advanced multimodal understanding. https://arxiv.org/abs/2412.10302.

[10] Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. arXiv preprint arXiv:2001.11314 (2020).

[11] 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. CoRR abs/2406.12793 (2024). https://doi.org/10.48550/arXiv.2406.12793