



† denotes multiple LLMs (e.g., GPT, Llama, Qwen).
‡ denotes multilingual prompts (e.g., English, Chinese, Japanese).