MMReD: A Benchmark for Multi-Modal Reasoning in Dense Contexts

Maxim Kurkin, Boris Shirokikh, Irina Abdullaeva, Viktoriia Chekalina

AIRI, Moscow, Russia
E-mail: kurkin@airi.net

Научный руководитель — *Andrey Kuznetsov*

Motivation. Current large language models (LLMs) and vision-language models (LVLMs) struggle with long-context reasoning, a key capability for applications like document analysis and video understanding. Existing benchmarks mainly focus on short contexts or retrieval-based tasks, failing to assess a model's ability to retain and reason over extended sequences. To address this, we propose MMReD (Multi-Modal Reasoning in Dense Context), a benchmark designed to systematically evaluate long-context reasoning for both text and vision modalities.

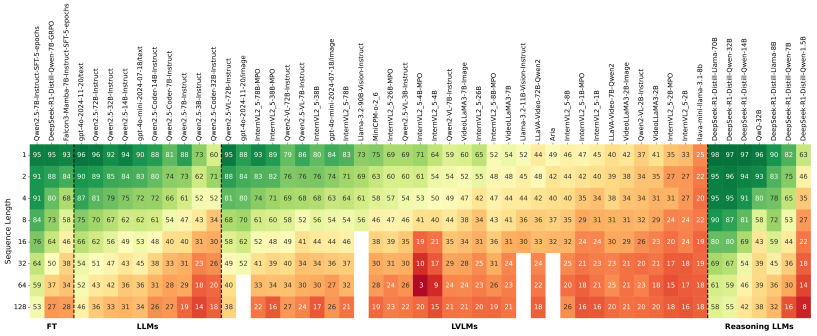
Benchmark Design. MMReD draws inspiration from bAbI [1] and includes two types of tasks: (1) Scene-Referenced (NIAH) tasks for extracting information from specific frames, and (2) Long-Context (LC) tasks for assessing comprehensive reasoning over entire sequences. The benchmark uses up to 128 frames with an exact-match accuracy metric. MMReD is designed to expose the limitations of current architectures and pretraining paradigms in handling long contexts effectively.

Key Findings. As shown in Figure 1, most models fail to maintain accuracy beyond 32 frames. Reasoning-specialized models like DeepSeek-R1 outperform standard LLMs and LVLMs, indicating that improved reasoning mechanisms might be more effective than simply increasing model size. The results also suggest that current pretraining methods are insufficient for long-context tasks.

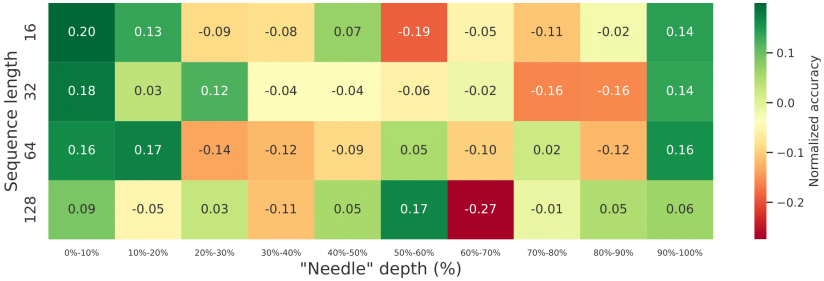
Lost-in-the-Middle Effect. Figure 2 shows no significant evidence of the lost-in-the-middle effect [2], suggesting that performance degradation is due to genuine long-context challenges rather than positional biases.

Architectural Analysis. Models using QFormer-based adapters and cross-attention mechanisms perform better on long-context tasks compared to those with simple MLP adapters. Video-oriented LVLMs, such as VideoLLaMA3, show improved performance due to advanced pooling techniques but still struggle with sequences longer than 64 frames.

Иллюстрации



Performance of different model families on MMReD across varying context lengths. Red indicates poor performance, green indicates good performance.



Impact of frame position on accuracy for scene-referenced tasks in MMReD. No significant drop in middle frames suggests that the lost-in-the-middle effect is minimal.

Литература

1. Weston, Jason et al. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks // 4th International Conference on Learning Representations (ICLR), 2016.
2. Bulatov, Aydar et al. Scaling Transformer to 1M tokens and beyond with RMT // arXiv preprint arXiv:2304.11062, 2023.