What Is That Talk About? A Video-to-Text Summarization Dataset for Scientific Presentations

Dongqi Liu[Ω], Chenxi Whitehouse[Δ], Xi Yu[Ω], Louis Mahon[Θ], Rohit Saxena[Θ], Zheng Zhao[Θ], Yifu Qiu[Θ], Mirella Lapata[Θ], Vera Demberg[ΩΨ]

[Ω]Saarland University, [Ψ]Max Planck Institute for Informatics
[Δ]University of Cambridge, [Θ]University of Edinburgh
dongqi.me@gmail.com

European Research Council
Established by the European Commission

ACL 2025 VIENNA

TL;DR

We present VISTA, the first scientific video-to-text summarization dataset, and show that plan-based method improves quality and factual accuracy over strong multimodal baselines

Motivation

- Why is scientific video-to-text summarization important?
- Why do existing large multimodal models (LMMs) struggle with scientific videos?
- What are the limitations of current summarization approaches?
- How does this paper address these gaps?

Motivation

- *Why is scientific video-to-text summarization important?*
  - Readers often prefer concise textual summaries to navigate dense video content
  - Unlike entertainment or news videos, scientific content demands factual precision and structured reasoning

Motivation

- *Why do existing LMMs struggle with scientific videos?*
    - Most LMMs are tuned for general-domain videos (YouTube, movies) — not technical talks
    - No large-scale, domain-specific benchmark has supported evaluation and adaptation of models in this setting

Motivation

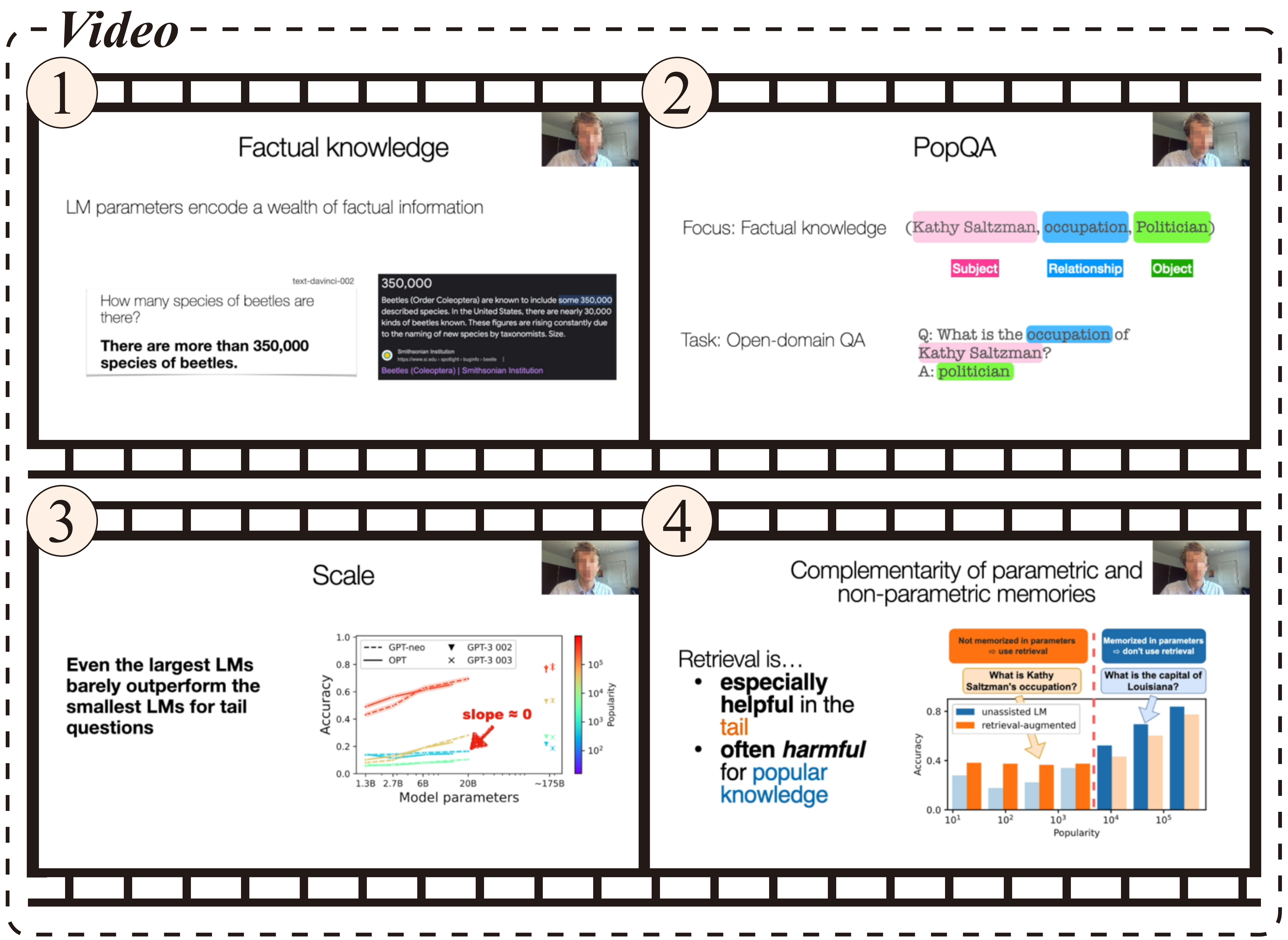
- *What are the limitations of current summarization approaches?*
    - SOTA LMMs show problems with structural grounding → incoherence, hallucination

Motivation

- *How does this paper address these gaps?*
  - VISTA (Video to Scientific Abstract) dataset
  - Planning method

What is VISTA?

- 18,599 AI conference presentation videos paired with corresponding paper abstracts
- Covers top-tier venues (ACL, NeurIPS, ICLR, etc.)



Summary

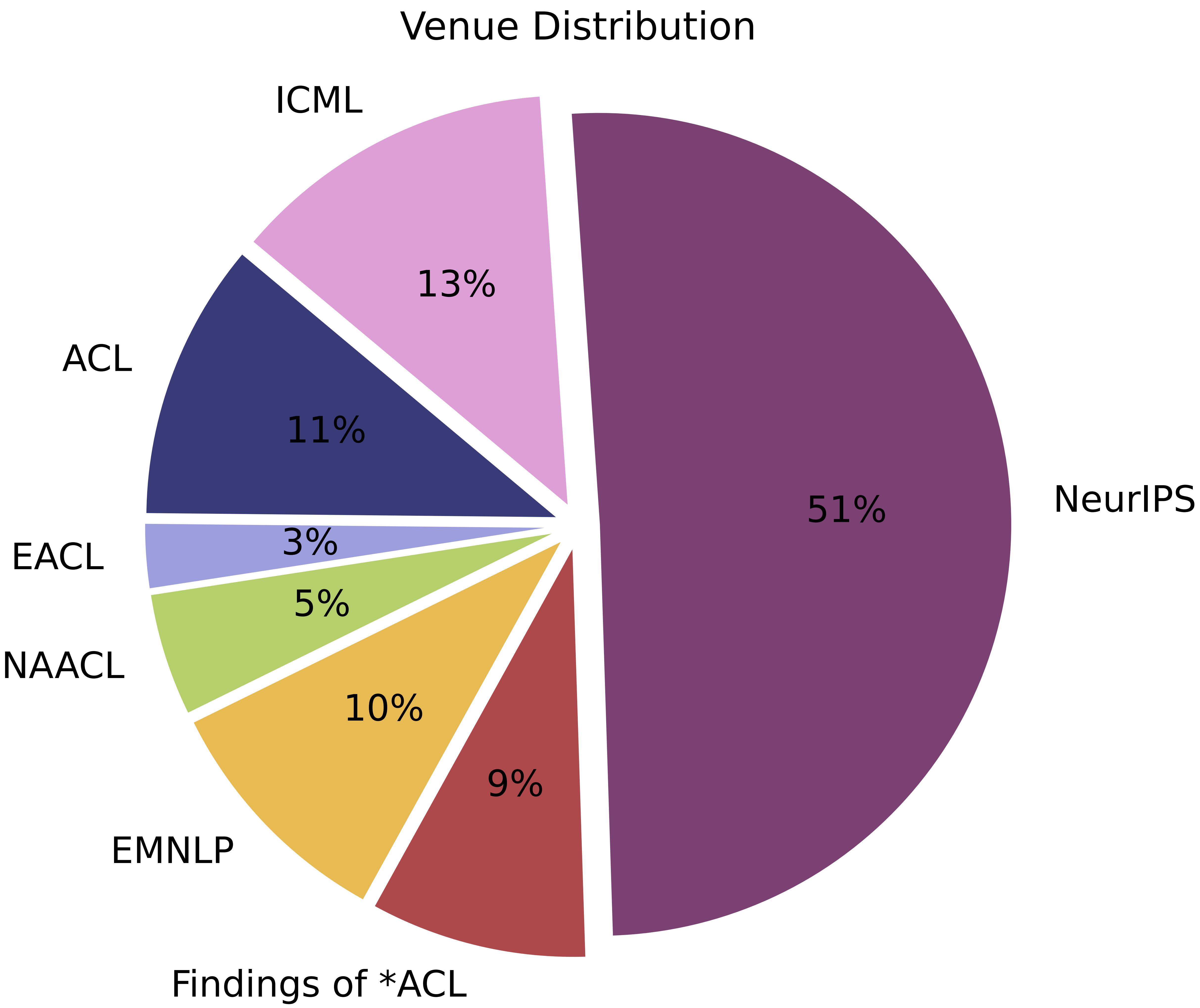
Despite their impressive performance on diverse tasks, large language models (LMs) [...], implying the difficulty of encoding a wealth of world knowledge in their parameters. This paper aims to understand LMs' strengths and [...], by [...]. We find that LMs struggle with less popular factual knowledge, and [...]. Scaling, on the other hand, mainly improves memorization of popular knowledge, and fails [...]. Based on those findings, we devise a new method for retrievalaugmentation[...] memories when necessary.

Quality Control

- Manual check: 500 random pairs reviewed by two PhDs — 0 rejections
- Automated check: GPT-o1 flagged 39 samples, all confirmed valid by manual review
- Quality criteria: *Each summary must be concise and accurately reflect the video content*

Dataset Split

- Data Split:
  - Training 80%, Validation 10%, Test 10%
- Filtering:
  - Only paper presentations (no tutorials/invited talks)
  - Videos: 1–30 min, English, 1-to-1 paper alignment



Dataset Statistics

- Most summaries remain under 250 tokens and 10 sentences
- Most videos last fewer than 10 minutes with under 30 shots

Dataset Comparison

- ❌ Existing datasets: short clips, casual topics (e.g., VideoXum, YouCook2, etc.)

- ❌ Prior datasets focus on narrations, actions, or subtitles

Dataset	Language	Domain	#Videos	VideoLen	SumLen
MSS (Li et al., 2017)	English, Chinese	News	50	3.4	—
YouCook2 (Zhou et al., 2018)	English	Cooking	2.0K	5.3	67.8
VideoStorytelling (Li et al., 2019)	English	Open	105	12.6	162.6
VMSMO (Li et al., 2020)	Chinese	Social Media	184.9K	1.0	11.2
MM-AVS (Fu et al., 2021)	English	News	2.2K	1.8	56.8
MLASK (Krubiński and Pecina, 2023)	Czech	News	41.2K	1.4	33.4
VideoXum (Lin et al., 2023)	English	Activities	14.0K	2.1	49.9
Shot2Story20K (Han et al., 2025)	English	Open	20.0K	0.3	201.8
BLiSS (He et al., 2023)	English	Livestream	13.3K	5.0	49.0
SummScreen^{3D} (Papalampidi and Lapata, 2023)	English	Open	4.5K	40.0	290.0
Ego4D-HCap (Islam et al., 2024)	English	Open	8.3K	28.5	25.6
Instruct-V2Xum (Hua et al., 2024)	English	Open	30.0K	3.1	239.0
MMSum (Qiu et al., 2024)	English	Open	5.1K	14.5	21.7
LfVS-T (Argaw et al., 2024)	English	YouTube	1.2K	12.2	—
VISTA (ours)	English	Academic	18.6K	6.8	192.6

Benchmarking

- Closed-source LMMs: GPT-o1, Gemini 2.0, Claude 3.5 Sonnet
- Open-source video LMMs: Video-LLaMA, Video-ChatGPT, Video-LLaVA, LLaVA-NeXT, mPLUG-Owl3
- Text baseline: LLaMA-3.1 (transcript, OCR)
- Audio baseline: Qwen2-Audio

Experiment Settings

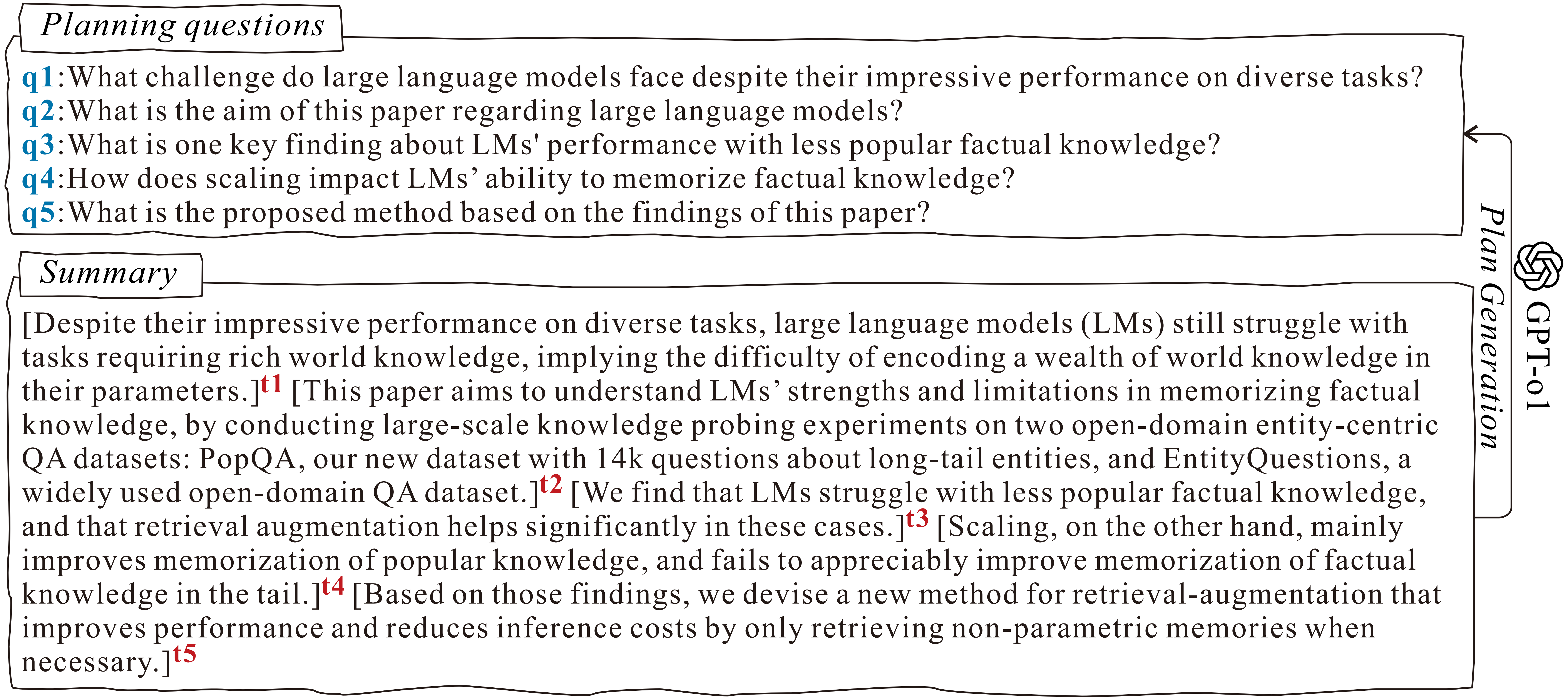
- Learning Settings:
  - Zero-shot inference, QLoRA fine-tuning, Full-parameter fine-tuning
- Training Details:
  - Standardized hyperparameters (AdamW, learning rate = 5e-5, batch size = 16, 16 epochs, early stopping)
- Video Preprocessing:
  - Video frames sampled at 0.1 fps, 32 frames per video
  - Transcription via Whisper, OCR via EasyOCR for text baselines

Evaluation Metrics

- Automated Metrics
  - ROUGE, SacreBLEU, METEOR, BERTScore, CIDEr-D
  - VideoScore: text–video alignment
  - FactVC: factual consistency
- Human Evaluation
  - 50 randomly sampled test videos
  - 3 expert annotators (double-blind)
  - Metrics: Faithfulness, Relevance, Informativeness, Conciseness, Coherence (Likert 1–5)

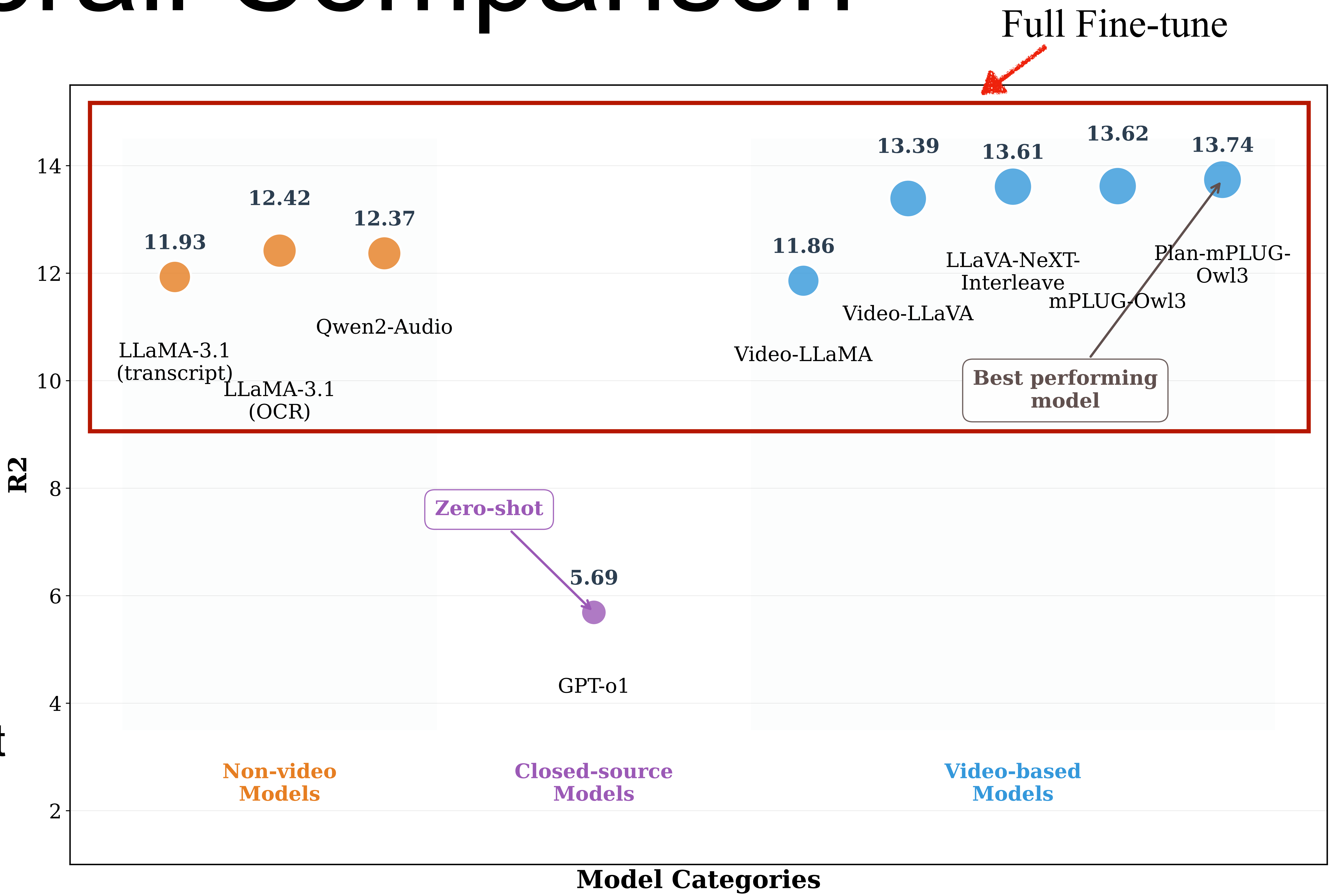
Plan-based Models

- Plan Generation (PG): generates question sequence
- Summary Generation (SG): generates summary answering plan questions



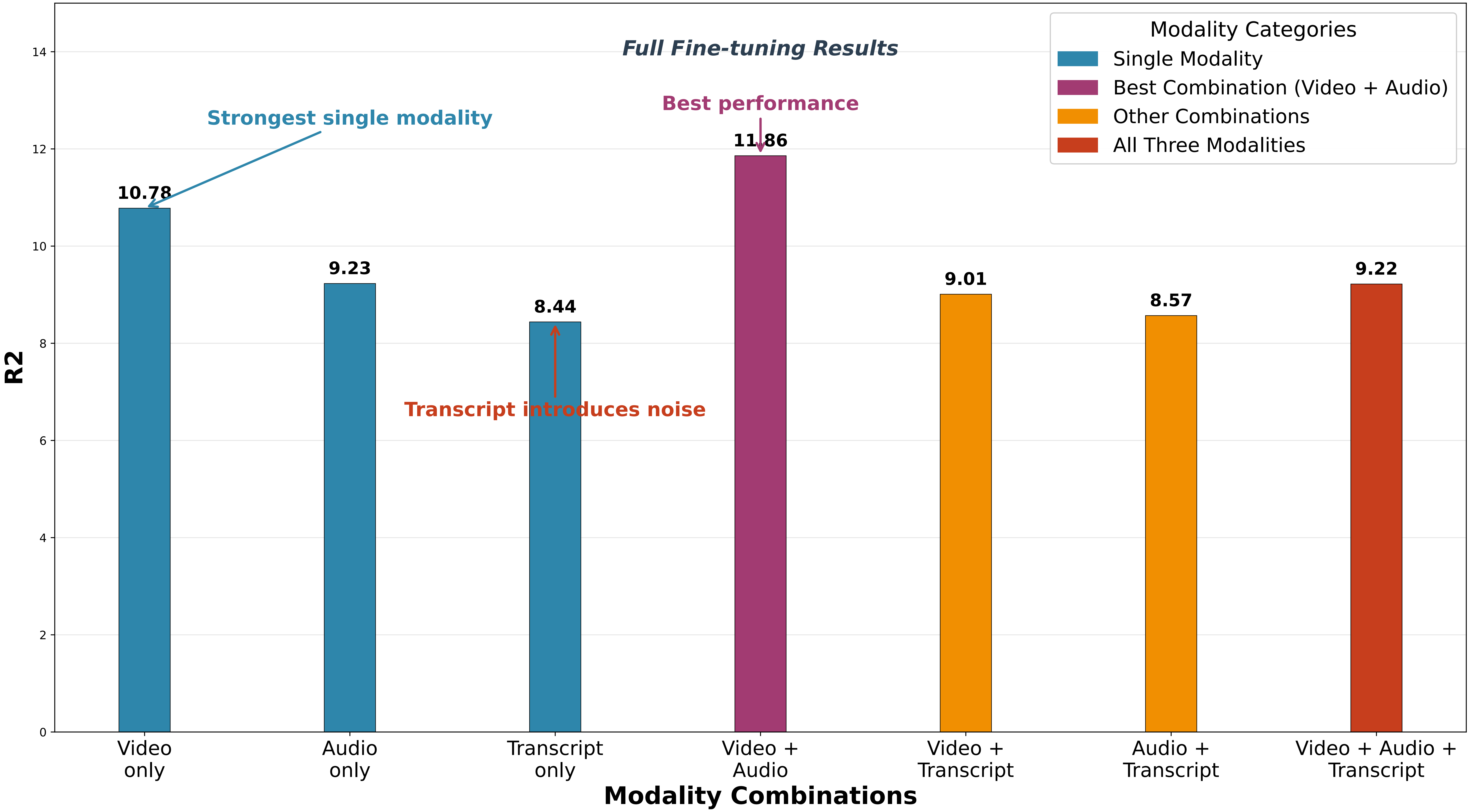
Overall Comparison

- Fine-tuning on in-domain VISTA data yields the largest gains
- Video-based LMMs **outperform** text- and audio-only models
- Closed-source models (e.g., GPT-o1) lead in zero-shot, but open-source models excel after fine-tuning
- Our plan-based method, built on mPLUG-Owl3 achieves **highest overall scores**



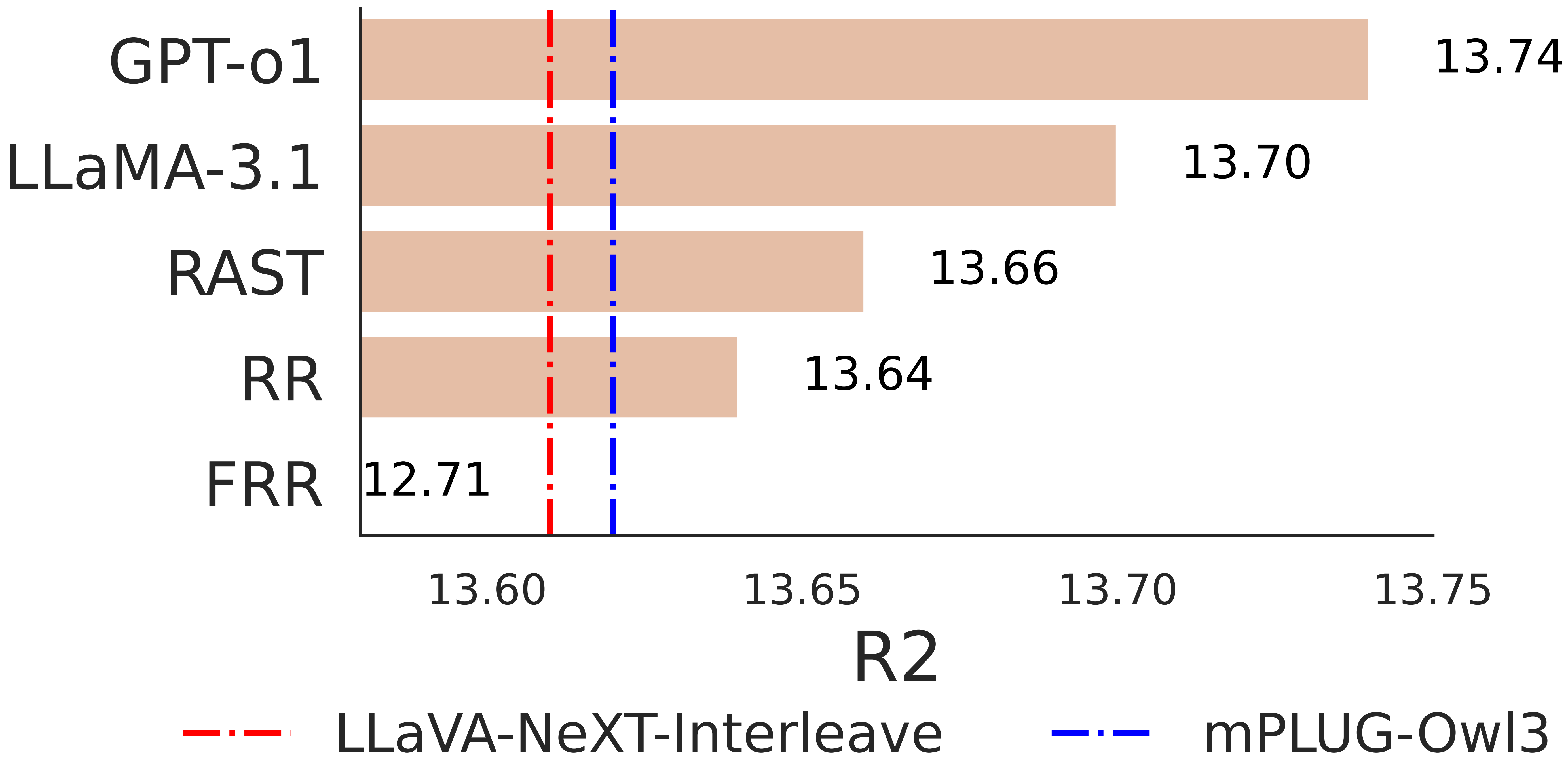
Modalities Matter

- Video is the strongest single modality (spatial-temporal cues)
- Adding audio provides minor gains; transcript (ASR) can introduce noise
- Best results from joint video + audio inputs



Impact of Planning Quality

- Higher-quality plans → better summaries
- Noise in planning (irrelevant/ random questions) **degrades** summary performance
- Plan-based approach remains robust under moderate noise

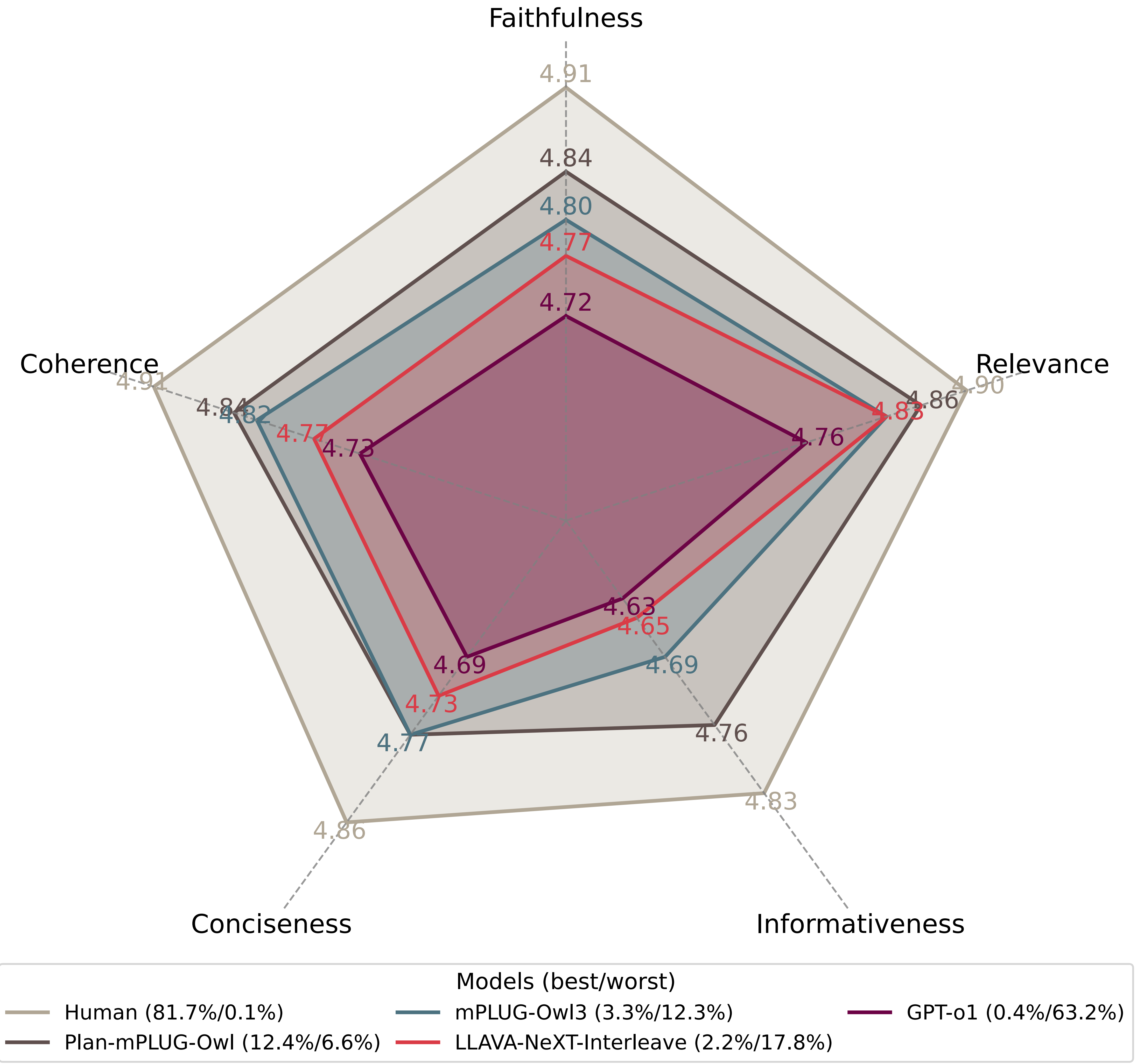


RAST (Gou et al., EMNLP 2023) is a SOTA question generation method.
RR = random replacement, FRR = full random replacement

Human Evaluation

- Human-written summaries **outperform** all neural LMMs on faithfulness, informativeness, coherence, etc.

- Plan-based model is **best among neural systems**, but gap with human remains



We compare with human performance, the top three finetuned models, and the best-performing closed-source model (under zero-shot setting).

Conclusion

- VISTA establishes a new benchmark for scientific video-to-text summarization
- Plan-based models improve summary quality and factual accuracy
- Significant gaps remain between model and human performance

More Info

- **Data & Code:** https://dongqi.me/projects/VISTA
- **Questions:** dongqi.me@gmail.com

Thanks for listening

Q&A

European Research Council
Established by the European Commission

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). Lapata acknowledges the support of the UK Engineering and Physical Sciences Research Council (Grant EP/W002876/1).