Findings of the Third Automatic Minuting (AutoMin) Challenge

Kartik Shinde
kartikkts46@gmail.com
Laurent Besacier
laurent.besacier@naverlabs.com
Ondřej Bojar
bojar@ufal.mff.cuni.cz
Thibaut Thonet
thibaut.thonet@naverlabs.com
Tirthankar Ghosal
ghosalt@ornl.gov

August 28th 2025

Introduction

- **AutoMin 2025**: Shared task on automatic meeting summarization into minutes.
- Task A: Summarize transcript into structured minutes.
- Task B **(New)**: Question answering (QA) based on meeting transcripts.
- Covered:
  - **two languages** (English, Czech), and
  - two domains (project meetings, EU Parliament).

Introduction

- Third edition of the challenge.
- Focus on structured meeting minutes.
- Extended with QA task.

Task A: Minuting

- Goal: Create structured minutes from transcripts.
- Languages: English, Czech.
- Domains: project meetings, EU Parliament sessions.
- Evaluation: GPT-4 LLM-as-a-judge.

Task B: Question Answering

- QA from meeting transcripts.
  - good opportunity to assess LLMs' ability to summarize and retrieve information from long contexts, equivalent to the length of an hour-long meeting
- Available only for project meetings domain.
- Mono- and cross-lingual settings.
- Evaluation: GPT-4o LLM-as-a-judge based on the reference answer and a numeric score rubric (Kim et al., 2024; Thonet et al., 2024) which explicits the expected quality criterion at each score level (1-10)

Datasets

- English + Czech transcripts.
- Project meetings (ELITR Minuting Corpus) and EU Parliament.
- Prepared for minutes and QA.

	English		Czech	
Meeting Minuted	#meetings	#hours	#meetings	#hours
Once	30	22	8	2
Twice	65	65	20	20
More than twice	25	22	31	31
Total meetings	120	109	59	53

Table: Basic transcript and minutes statistics for ELITR Minuting Corpus.

New task: QA from meeting transcripts

Q: Who were the participants of the meeting?
A: [PERSON14], [PERSON10], [PERSON5], [PERSON9], [PERSON1], [PERSON11]

Q: What was the main purpose of this meeting?
A: Discuss and finalize the technical setup for a demo

Q (Conv version): How many scenarios were discussed?
Q (QA version): How many scenarios were discussed for the demo setup?
A: 3 (plans A, B and C)
Q (Conv version): Which scenario was chosen eventually?
Q (QA version): Which scenario was chosen eventually for the demo setup?
A: Plan C

Figure: An illustration of questions for the QA track (QA version was used, Conv not used)

Task A Submissions

1 participant only (lower than in previous years)

HallucinationIndexes:

- RL-based approach to reduce hallucinations in summarization.
- Introduces Entity Hallucination Index (EHI) as reward signal.
- Model-centric, domain-agnostic, fine-tuned BART for factual consistency.
- Does not rely on meeting-specific preprocessing (no speaker segmentation or dialogue structuring).

GPT-2025 Baseline:

- GPT-4 baseline using same prompt as in 2023: "Summarize the following project meeting in 5–10 bullet points."

Task B Submissions

2 participants' submissions plus baselines using LLMs

GETALP:

- Combines Retrieval Augmented Generation (RAG) and Abstract Meaning Representation (AMR).
- Three systems: RAG-only, AMR-only, RAG+AMR.
- Extractive: prompted LLaMA-3.1-8B selects 1–2 most relevant sentences.
- Evaluated on monolingual (EN) and cross-lingual (EN–CZ) tasks.

HallucinationIndexes: Same RL-based approach as Task A, base model Flan-T5 (1024-token context window), focusing on entity-level faithfulness.

Baselines: GPT-4o, LLaMA-3.2-3B, LLaMA-3.1-8B, Phi-4-Mini, Phi-3-Small-128k.

Task A (Minuting) Evaluation (1/2)

We consider the criteria from Ghosal et al. (2023):

1. **Adequacy** assesses if the minutes adequately capture the major topics discussed in the meeting, also considering coverage (all such topics covered).
2. **Fluency** reflects if the minutes consist of fluent, coherent texts and are readable to the evaluator.
3. **Grammatical Correctness** checks the level to which the minutes are grammatically correct.
4. **Relevance** signifies the extent to which the minutes overall capture the important content from the source transcript (as opposed to summarizing useless parts).

. . . but. . .

- No funds for manual evaluation this year, only 1 participant. . .

$\Rightarrow$ Examine also 2023 systems and experiment with LLM-based evaluation.

Task A (Minuting) Evaluation (2/2)

- GPT-4 improved internally.
- We improved the prompt (Chain-of-Thought incl. "reflection").

⇒ We have 3 LLM-derived evaluations to consider:

GPT-2025-CoT: scores collected this year with the current GPT and the Chain-of-Thought prompt. "Function calls" help with formal format and output extraction.

GPT-2023-AFGR scores collected in 2023 with GPT back then, and the simple prompt asking for Adequacy, Fluency, Grammaticality and Relevance (AFGR) in one go.

GPT-2025-AFGR the current GPT using the simple AFGR prompt.

Results: Task A (Minuting): GPT-2025-CoT ELMI EN

ELMI EN	Adeq.	Flu.	Gram.	Relev.
GPT-4	**4.00±0.43**	**4.42±0.51**	**4.75±0.45**	**4.33±0.65**
text-davinci-003	3.25±0.62	3.92±0.67	4.00±0.74	3.33±0.78
davinci-003	3.08±0.51	3.75±0.45	≀4.08±0.29	3.25±0.87
Zoom-long	2.42±0.51	3.00±0.60	3.25±0.75	2.42±0.51
Synapse	2.33±0.49	≀3.25±0.45	3.17±0.58	2.42±0.51
Darbarer	2.25±0.45	≀3.50±0.52	≀3.83±0.58	2.33±0.49
Zoom-short	2.17±0.39	3.00±0.60	3.67±0.49	2.17±0.39
Team Iterate	2.08±0.51	2.50±0.52	3.00±0.43	2.08±0.51
NTR	2.00±0.00	2.08±0.29	2.17±0.39	2.00±0.00
GPT-4 (2025)	**3.67±0.49**	**4.42±0.51**	**4.75±0.45**	**4.00±0.60**
HallucinationIndexes@ (2025)				
bart-samsum	1.08±0.29	2.92±0.51	3.50±0.67	1.17±0.39
distilbart	1.08±0.29	2.17±0.39	2.25±0.45	1.08±0.29
pegasus-xsum	1.00±0.00	2.17±0.58	≀2.92±0.90	1.00±0.00
t5-small	1.00±0.00	2.08±0.51	2.42±0.67	1.00±0.00

Results: Task A (Minuting): GPT-2025-CoT ELMI EN

ELMI EN	Adeq.	Flu.	Gram.	Relev.
GPT-4	**4.00±0.43**	**4.42±0.51**	**4.75±0.45**	**4.33±0.65**
text-davinci-003	3.25±0.62	3.92±0.67	4.00±0.74	3.33±0.78
davinci-003	3.08±0.51	3.75±0.45	4.08±0.29	3.25±0.87
Zoom-long	2.42±0.51	3.00±0.60	3.25±0.75	2.42±0.51
Synapse	2.33±0.49	3.25±0.45	3.17±0.58	2.42±0.51
Darbarer	[illegible]	3.50±0.52	3.83±0.58	2.33±0.49
Zoom-short	[illegible]	3.00±0.60	3.67±0.49	2.17±0.39
Team Iterate	[illegible]	2.50±0.52	3.00±0.43	2.08±0.51
NTR	[illegible]	2.08±0.29	2.17±0.39	2.00±0.00
GPT-4 (2025)	**3.67±0.49**	**4.42±0.51**	**4.75±0.45**	**4.00±0.60**
HallucinationIndexes@ (2025)				
bart-samsum	1.08±0.29	2.92±0.51	3.50±0.67	1.17±0.39
distilbart	1.08±0.29	2.17±0.39	2.25±0.45	1.08±0.29
pegasus-xsum	1.00±0.00	2.17±0.58	2.92±0.90	1.00±0.00
t5-small	1.00±0.00	2.08±0.51	2.42±0.67	1.00±0.00

Unfortunately, all variants from our participant are rather poor.

Results: Task A (Minuting): GPT-2025-CoT ELMI EN

ELMI EN	Adeq.	Flu.	Gram.	Relev.
GPT-4	**4.00±0.43**	**4.42±0.51**	**4.75±0.45**	**4.33±0.65**
text-davinci-003	3.25±0.62	3.92±0.67	4.00±0.74	3.33±0.78
davinci-003	3.08±0.51	3.75±0.45	≀4.08±0.29	3.25±0.87
Zoom-long	2.42±0.51	3.00±0.60	3.25±0.75	2.42±0.51
Synapse	2.33±0.49	≀3.25±0.45	3.17±0.58	2.42±0.51
Darbarer		2	≀3.83±0.58	2.33±0.49
Zoom-short		0	3.67±0.49	2.17±0.39
Team Iterate		2	3.00±0.43	2.08±0.51
NTR		2.08±0.29	2.17±0.39	2.00±0.00
GPT-4 (2025)	**3.67±0.49**	**4.42±0.51**	**4.75±0.45**	**4.00±0.60**
HallucinationIndexes@ (2025)				
bart-samsum	1.08±0.29	2.92±0.51	3.50±0.67	1.17±0.39
distilbart	1.08±0.29	2.17±0.39	2.25±0.45	1.08±0.29
pegasus-xsum	1.00±0.00	2.17±0.58	≀2.92±0.90	1.00±0.00
t5-small	1.00±0.00	2.08±0.51	2.42±0.67	1.00±0.00

ELMI EN GPT-4 summaries are deemed a little worse than GPT-4-2023 summaries (in GPT-4-2025 evaluation).

Results: Task A (Minuting): GPT-2025-CoT EuroParlMin

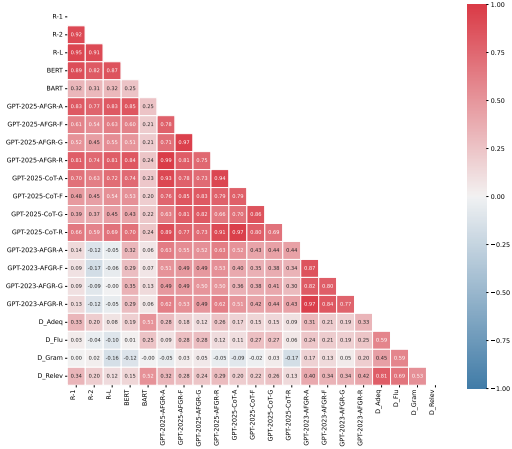
EuroParlMin	**Adeq.**	**Flu.**	**Gram.**	**Relev.**
Synapse	**3.40±0.80**	**4.62±0.58**	**4.83±0.51**	**3.99±0.78**
NTR	2.80±1.04	3.80±1.01	4.02±1.07	3.10±1.22
Darbarer	2.43±0.69	≀4.29±0.64	≀4.85±0.44	3.02±0.83
GPT-4	1.58±1.24	3.90±0.68	4.70±0.57	1.64±1.33
davinci-003	1.57±1.21	3.84±0.65	4.64±0.54	1.64±1.32
GPT-4 (2025)	**4.78±0.58**	**4.95±0.27**	**4.98±0.26**	**4.84±0.51**
HallucinationIndexes@ (2025)				
bart-samsum	2.96±0.99	4.41±0.68	4.63±0.76	3.52±1.16
distilbart	2.73±0.96	4.02±0.85	4.32±1.00	3.23±1.16
t5-small	2.67±0.98	3.00±0.69	3.09±0.87	3.22±1.27
pegasus-xsum	1.54±0.63	≀3.31±1.03	≀4.02±1.21	1.84±0.79

Results: Task A (Minuting): GPT-2025-CoT EuroParlMin

EuroParlMin	Adeq.	Flu.	Gram.	Relev.
Synapse	**3.40±0.80**	**4.62±0.58**	**4.83±0.51**	**3.99±0.78**
NTR		01	4.02±1.07	3.10±1.22
Darbarer		64	≀4.85±0.44	3.02±0.83
GPT-4		58	4.70±0.57	1.64±1.33
davinci-003		3.64±0.65	4.64±0.54	1.64±1.32
GPT-4 (2025)	**4.78±0.58**	**4.95±0.27**	**4.98±0.26**	**4.84±0.51**
HallucinationIndexes@ (2025)				
bart-samsum	2.96±0.99	4.41±0.68	4.63±0.76	3.52±1.16
distilbart	2.73±0.96	4.02±0.85	4.32±1.00	3.23±1.16
t5-small	2.67±0.98	3.00±0.69	3.09±0.87	3.22±1.27
pegasus-xsum	1.54±0.63	≀3.31±1.03	≀4.02±1.21	1.84±0.79

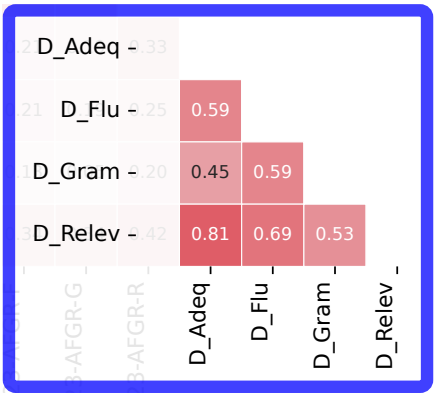
On EuroParlMin, GPT-4 summaries are deemed better this year than in 2023 (in GPT-4-2025 evaluation).

Correlations of Diverse Evaluation Methods



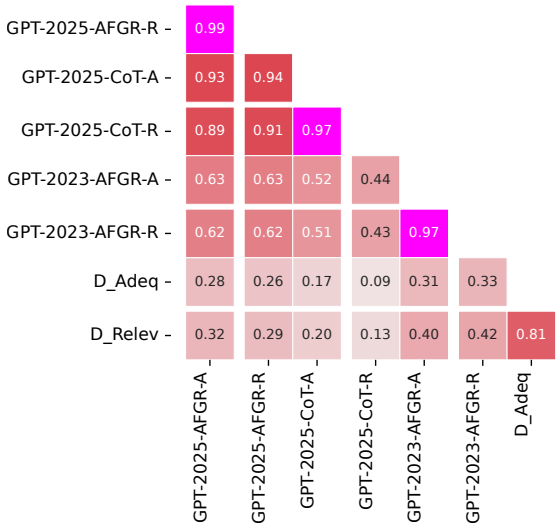
- Pearson correlations of scores by different scoring methods across *individual meetings*.

Manual Scores (2023 data): Relevance ~ Adequacy at .8 Pearson

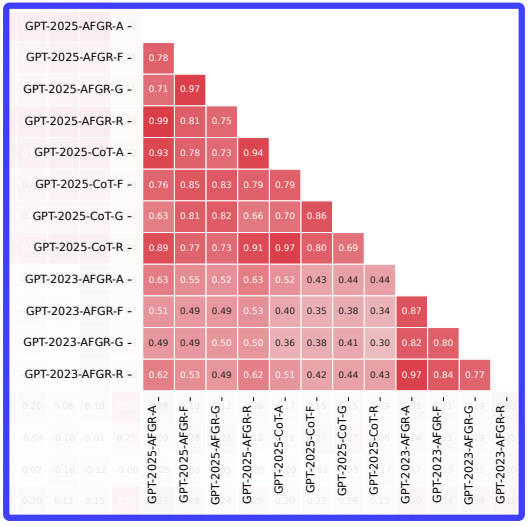


- Adequacy: recall-like: "Do the minutes summarize all important things?"
- Relevance: precision-like: "Are relevant parts summarized (as opposed to useless ones)?"

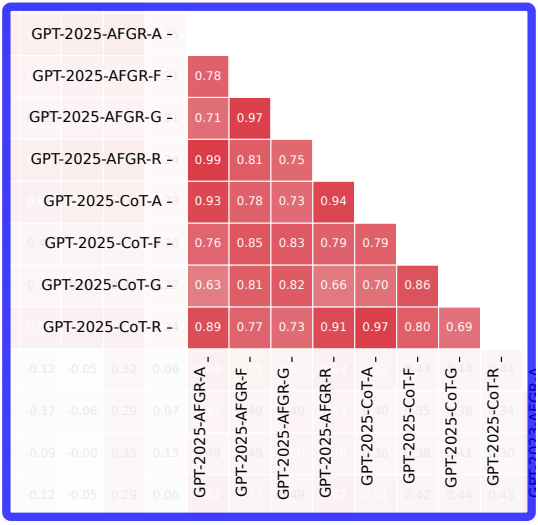
Relevance ~ Adequacy Even More for GPT



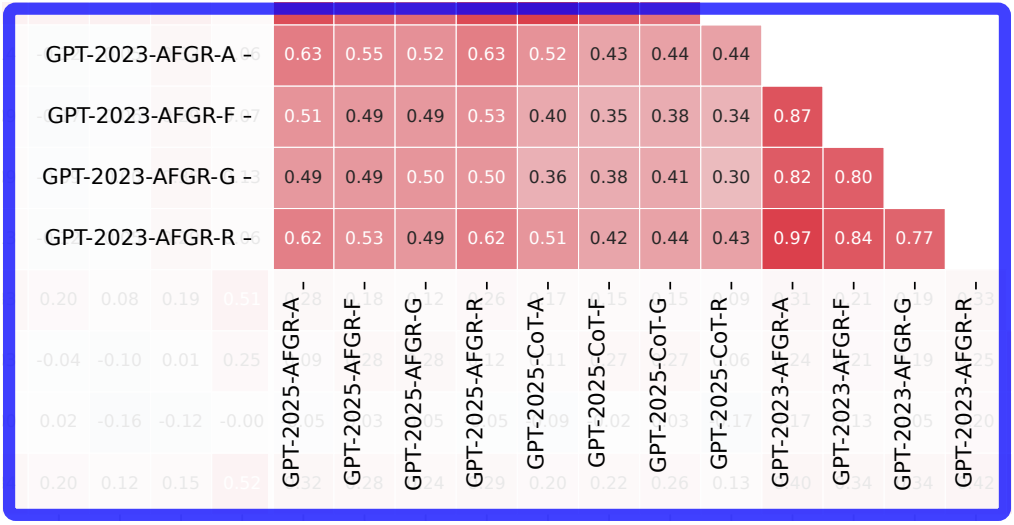
All GPT Variants Correlated from ~.3 to .99



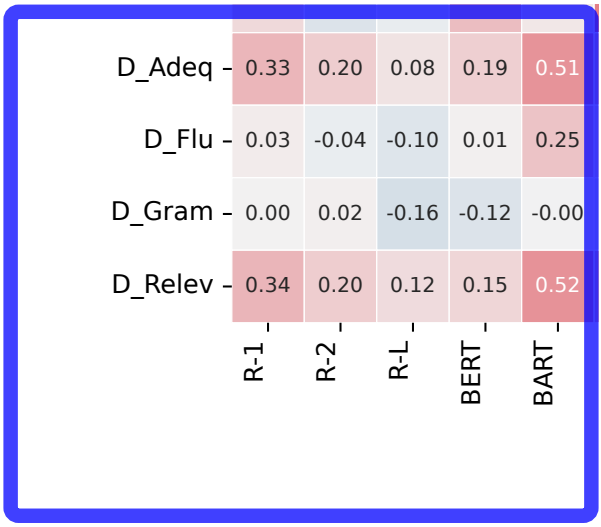
Within GPT-2025-eval, CoT ~ AFGR; from ~.6 to .99



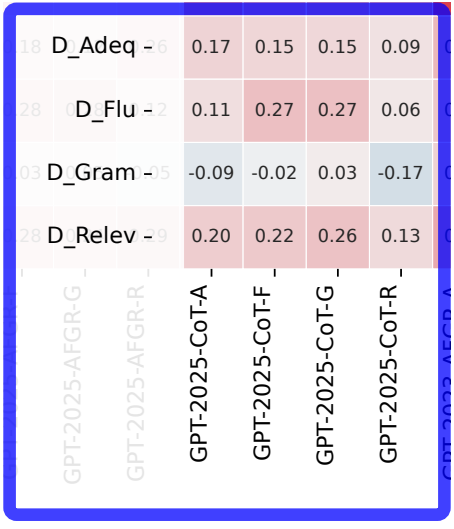
Within GPT-eval, 2023 $\not\approx$ 2025, while A~F~G~R



Classical Evaluation vs. Manual: .3 Max



Manual vs. GPT-2025-CoT: .27 Max ☹, Adequacy at .15 ☹



A Small Adversarial Study in Task A

- GPT-4 outputs from 2023 were messed up for EuroParlMin.
- We manually revised 91 meetings, finding 45 misaligned minutes.

$\Rightarrow$ A transcript with *wrong minutes should score low* in evaluation.

Average Metric Scores for Good and Adversary Minutes

	Adversarial Minutes	Regular Minutes	Δ	Sign. Diff
GPT-2025-CoT-R	1.24±0.77	4.68±0.61	3.44	✓ (p=0.0000)
GPT-2025-CoT-A	1.20±0.59	4.61±0.62	3.41	✓ (p=0.0000)
GPT-2025-AFGR-R	1.37±0.54	4.19±0.60	2.82	✓ (p=0.0000)
GPT-2025-AFGR-A	1.43±0.51	4.22±0.60	2.79	✓ (p=0.0000)
BART	-4.75±0.61	-3.97±0.49	0.78	✓ (p=0.0000)
GPT-2025-AFGR-F	3.99±0.72	4.64±0.32	0.65	✓ (p=0.0000)
GPT-2025-CoT-F	4.64±0.91	4.93±0.25	0.29	✓ (p=0.0071)
GPT-2025-AFGR-G	4.72±0.88	4.95±0.14	0.23	✗ (p=0.2631)
GPT-2025-CoT-G	4.82±0.83	4.98±0.15	0.16	✗ (p=0.4069)
BERT	0.12±0.12	0.26±0.13	0.14	✓ (p=0.0000)
R-1	0.17±0.06	0.30±0.13	0.13	✓ (p=0.0000)
R-2	0.03±0.02	0.12±0.08	0.09	✓ (p=0.0000)
R-L	0.11±0.04	0.19±0.09	0.08	✓ (p=0.0000)

- GPT-2025-CoT-R best in scoring adversary minutes low.
- Even classical metrics notice a significant (✓) difference (Mann-Whitney).

Results: Task B (QA)

Monolingual (EN)

Approach	Mean Score
baseline_gpt-4o-2024-11-20	7.74
baseline_llama-3.1-8B-instruct	7.08
baseline_phi-4-mini-instruct	6.84
baseline_phi-3-small-128k-instruct	6.65
baseline_llama-3.2-3B-instruct	6.33
GETALP@AutoMin	4.55
GETALP@AutoMin_amr	4.34
GETALP@AutoMin_amr_only	2.73
HallucinationIndexes@AutoMin	2.28

Cross-lingual (EN-CZ)

Approach	Mean Score
baseline_gpt-4o-2024-11-20	7.69
baseline_llama-3.1-8B-instruct	6.21
baseline_phi-4-mini-instruct	5.41
baseline_llama-3.2-3B-instruct	5.11
baseline_phi-3-small-128k-instruct	4.77
GETALP@AutoMin_amr	3.11
GETALP@AutoMin	2.85
GETALP@AutoMin_amr_only	2.15

Task B: Comments on Results

- **Monolingual (EN) observations:**
  - GPT-4o and LLaMA-3.1-8B perform strongly on long-context meeting QA, closed models outperform open ones.
  - Phi-4-Mini-Instruct baseline performs competitively (strong 3–4B parameter model).
  - Participant submissions (*GETALP*, *HallucinationIndexes*) score lower than baselines.
  - Likely due to extractive approaches being disadvantaged by LLM-based evaluation; less effective for complex, nuanced questions.
  - *HallucinationIndexes* limited by Flan-T5 context window (1,024 tokens) vs. long transcripts ($\geq$ 10,000 tokens).
- **Cross-lingual (EN–CZ) observations:**
  - GPT-4o remains robust, close to monolingual performance.
  - Open models (LLaMA, Phi) show more than 1 point drop in score.
  - Only *GETALP* submitted a cross-lingual system; performance notably lower than their monolingual results.

Conclusion

- AutoMin 2025 advanced meeting summarization research.
- Introduced QA as new track.
- GPT-based evaluation in Task A:
  - Little difference between simple vs. Chain-of-Thought.
  - Larger difference between 2025 run vs. 2023 run (same simple prompt).
  - Adequacy $\approx$ Relevance.
- Minuting evaluation is *far from solved*.
  - Classical evaluation methods (ROUGE) not very good (.3 Pearson for A and R).
  - GPT-based evaluation methods even slighly worse (.15 Pearson for A and R).
  - At least, both can identify mixed-up transcript and minutes.

$\Rightarrow$ Cannot really recommend LLM-as-a-Judge.

- Future of meeting processing in the age of generalist LLMs?

Acknowledgments

Thanks to all participants and organizers.

This paper was partially funded by the European Commission through the UTTER project under grant number 101070631.

This work has also received funding from the Project OP JAK Mezisektorová spolupráce Nr. CZ.02.01.01/00/23_020/0008518 named "Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím."

References I

Tirthankar Ghosal, Ondřej Bojar, Marie Hledíková, Tom Kocmi, and Anna Nedoluzhko. 2023. Overview of the Second Shared Task on Automatic Minuting (AutoMin) at INLG 2023. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 138–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing evaluation capability in language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Thibaut Thonet, Jos Rozen, and Laurent Besacier. 2024. Elitr-bench: A meeting assistant benchmark for long-context language models. *Preprint*, arXiv:2403.20262.