会議有効性の再考：時間的細粒度自動会議評価向けのベンチマークとフレームワーク

Yihang Li　Chenhui Chu
京都大学大学院情報学研究科
{liyh, chu}@nlp.ist.i.kyoto-u.ac.jp

概要

会議の有効性を評価することは、組織の生産性を向上させる上で重要である。従来手法は参加者への事後アンケートに依存しており、会議全体に対して単一のスコアを付与することにとどまっていた。このような人手評価への依存は、スケーラビリティ、コスト、再現性の観点から限界がある。また、単一スコアでは議論の動的な性質を捉えきれない。そこで本研究では、時間的細粒度に着目した会議有効性評価の新たなパラダイムを提案する。有効性を「単位時間あたりの目標達成率」と定義し、会議をトピックごとのセグメントに分割して評価を行う。本研究のために、AMI Corpus の 130 会議から抽出した 2,460 セグメントに対し、人手による有効性スコアを付与したメタ評価データセット「AMI-ME」を構築した。さらに、大規模言語モデルを用いて、会議全体の目標に対する各セグメントの有効性を自動評価するフレームワークを提案する。実験により、本フレームワークの有効性を示し、新たなベンチマークを確立した。

1　はじめに

会議は共同作業に不可欠である一方で、非効率の原因として挙げられることも多い [1, 2, 3, 4, 5]。会議の生産性を向上させるには会議の有効性を正確に評価する必要があるが、評価手法に関する統一的な見解は未だに得られていない [6]。既存の評価手法は、主に参加者への事後アンケートに依存しており、会議全体に対して粗い粒度の単一スコアを与えるのみである [6, 7, 8, 9, 10, 11, 12]。これは会議の事後的な印象に過ぎず、議論の動的な推移や、生産的な時間帯とそうでない時間帯の区別を捉えきれない。加えて、人手評価はコストが高く、規模化や再現性の確保が困難である。また、会議データはプライバシーの問題を含み、収集が難しいため、データ不足も研究のボトルネックとなっていた。

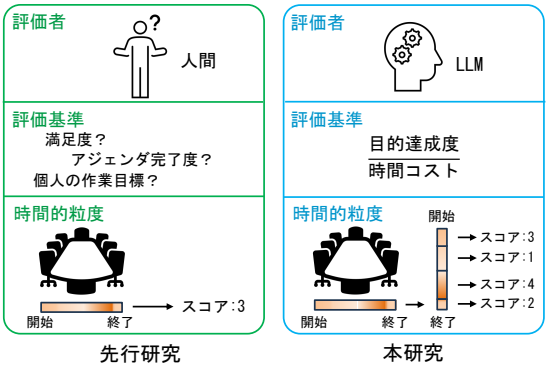


図 1: 会議有効性評価のパラダイム

そこで本研究では、図 1 に示すように、2 つの原則に基づく新たな評価パラダイムを提案する。第一に、有効性を「単位時間あたりの目標達成率」と定義し、客観的かつ普遍的な評価基準を導入する。第二に、時間的細粒度評価手法を提案する。会議全体に単一のスコアを割り当てるのではなく、会議をトピックセグメントに分割し、セグメントごとの有効性を評価する。このアプローチにより、会議のダイナミクスの詳細な分析が可能になり、評価の精度が向上するとともに、モデルの学習・評価に利用可能なデータの規模を大幅に増やせる。

このパラダイムに基づき、本研究では、模擬ビジネス会議データセットである AMI Corpus [13] をベースに会議のメタ評価データセット「AMI-ME」を構築した。まず、AMI Corpus の元のセグメンテーションを連続的な細粒度の単位に修正した。そして、各セグメントの有効性をスコアリングする包括的な人手アノテーションを実施した結果、130 会議から 2,460 のセグメントとそれに対応する有効性スコアが得られた。

さらに、会議の有効性を評価する LLM ベースのフレームワークを提案する。フレームワークではトランスクリプトに対してトピック分割を行い、会議

全体の目標に対する貢献度に基づいて各セグメントの有効性を評価する。実験でフレームワークの有効性を検証し、会議分析および対話エージェントにおける将来の研究のためのベンチマークを確立した。

2 関連研究

会議に対する研究自体が進展したものの、会議の有効性を評価する方法についての合意は形成されていない。先行研究の多くは、会議全体に対して単一のスコアを導出するために、参加者への事後アンケートに依存している。評価基準は研究によって大きく異なっている。既存の文献調査を通じて、これまでの基準は主に 2 つのタイプに分類できる。1 つ目は客観的基準であり、これはあらゆる観察者がアクセス可能な情報に基づいており、個人に依存しないように設計されている。目標達成度 [7, 9, 10]、アジェンダ完了度 [8] などが客観的基準に該当し、異なる評価者間での一貫した評価を目指している。2 つ目は主観的基準であり、参加者に個人の感情や文脈に基づいて評価を求めるものであり、評価者個人に依存する。個人の作業目標達成 [4, 12] や決定への満足度 [7] 等が主観的基準として挙げられるが、会議の観察可能な文脈のみによる評価は困難である。

3 評価基準と時間的細粒度評価

主観的評価の難しさを考慮し、本研究では客観的な評価基準を採用する。具体的には、会議の目標達成度と時間コストの 2 つの要素に着目し、有効性を以下のように定義する：会議の有効性 = 目標達成度 / 時間コスト。この式は、コスト（参加者が費やした時間）に対する成果（共通目標の達成）の比率という「効率」の基本的定義と整合する。ここで「目標達成度」は、各目標を個別に評価するのではなく、すべての目標に対する集団的な進捗の全体的な尺度であり、各目標の相対的な重要性を暗黙的に重み付けしている。また、「会議の目標」は、事前に計画されたものに限らず、会議終了時にその内容から統合された創発的な目標も含むものとして定義する。

この定義に基づき、本研究では会議全体に単一スコアを付与する従来手法ではなく、会議をトピックごとのセグメントに分割し、それぞれを評価する時間的細粒度評価手法を提案する。この手法には主に 3 つの利点がある。第一に、会議内の効率的な期間と非効率な期間の差異を捉え、詳細な分析が可能になる点である。第二に、短く焦点の絞られたセグメントの方がアノテーターにとって評価しやすいため、アノテーション精度の向上が期待できる。第三に、各会議を分割することでデータ数を大幅に増やし、データの希少性を克服して頑健な統計分析を促進できる点である。本研究では、このアプローチを実現するために、議論のトピックに基づいて会議を分割する手法を採用する。

4 データセット構築手順

本節では、会議有効性メタ評価データセット AMI-ME の構築について説明する。手順は、参照ベースのトピック分割、およびセグメント有効性の人手アノテーションからなる。

4.1 参照ベースのトピック分割

本研究では、主に模擬ビジネス会議で構成され、専門的なドメイン知識なくアノテーターが理解可能な AMI Corpus [13] を基にデータセットを構築した。AMI Corpus から 130 の会議を選定した。各会議は通常参加者 4 名で構成され、平均時間は 31 分である。

AMI Corpus にはトピック境界のアノテーションが存在するが、不連続であったり粒度が粗かったりする場合がある。これらの問題に対処するために、参照ベースのトピック分割手法を導入した。この手法では、元の AMI のアノテーションを参照として、LLM を使用して連続的かつより細かい粒度の分割を生成する。得られたセグメント数およびセグメント時間の分布を、それぞれ付録 A 図 3(a) と 3(b) に示す 。比較分析の結果、このタスクには Gemini-2.5-Pro を選定した。人手による確認の結果、一貫した議論の途中に境界を設けるなどの重大なエラーはほとんど見られないことがわかった。付録 A 図 3(c) に示すように、結果として得られた分割は、元の 2,109 箇所の境界のうち 1,668 箇所を包含し、新たに 661 箇所を導入した一方で、連続性を確保するために多くの元の境界を除外している 。

4.2 有効性に対する人手アノテーション

トピック分割後、セグメントの有効性について人手によるアノテーションを行った。アノテーション画面（付録図**??**を参照）では、評価基準、トピックで分割されたトランスクリプト、人手によって事前に定義された会議目標のセット、および会議の理解を助けるための人手作成または LLM が生成した補助情報を含む、豊富な文脈を提供した。さらに、各

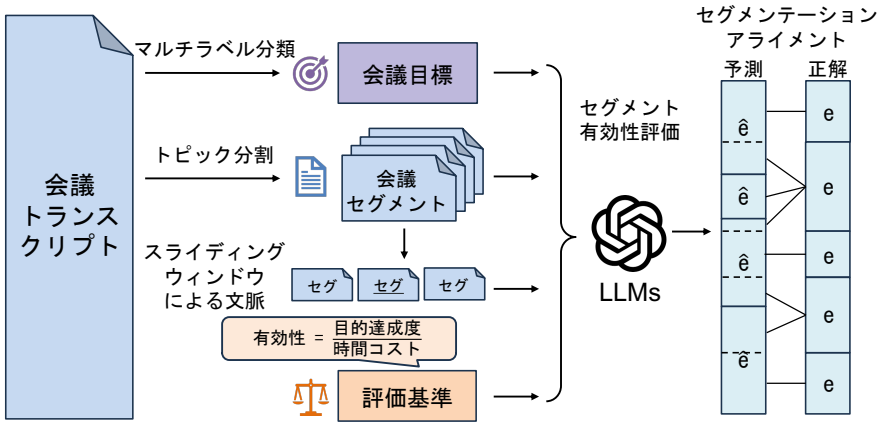


図 2: 自動評価フレームワーク

発話がタイムスタンプ付きで表示され、アノテーターが有効性を評価する際にセグメントの時間コストを考慮できるようにした。各セグメントについて、アノテーターはそのセグメントが寄与した目標を特定し、5 段階評価で有効性スコアを付与した。各会議セグメントは、3 人のアノテーターによって独立して評価された。

4.3 データセットの分析

構築された AMI-ME データセットには、130 会議からの 2,460 セグメントが含まれており、それぞれに 3 つの有効性スコアとマルチラベル形式での目標アノテーションが付与されている。有効性スコアの分布を付録 A 図 3(d) に示す 。アノテーション対象を 63 会議と 67 会議の 2 つのグループに分けて、アノテーター間の一致度を級内相関係数 ICC(2, k) [14] で評価した結果、スコアはそれぞれ 0.8769 および 0.8202 となり、共に「良好 (Good)」な水準の信頼性が確認された。

5 自動評価フレームワーク

図 2 に示すように、LLM による会議の有効性を自動評価するフレームワークを提案する。このフレームワークは、会議全体の目標の特定、トピック分割の実行、およびセグメントの有効性評価を行う。予測された分割と正解の分割の不一致を考慮し、メタ評価をサポートするためのセグメンテーションアライメント手順も実装している。

5.1 セグメント有効性評価

セグメントの有効性を評価するために、評価ステップの生成に Chain-of-Thought を用いたフォーム入力パラダイムを使用する G-Eval [15] を採用した。モデルへの入力は、異なるレベルの有効性を定義する詳細な評価基準、会議目標、ターゲットセグメントのトランスクリプト、およびその周囲のコンテキストセグメントから構成される。これらの入力に基づき、モデルはセグメントが時間を効率的に利用しつつ会議全体の目標にどの程度効果的に貢献したかを評価し、対応する有効性スコアを割り当てる。

会議の目標はマルチラベルアプローチを使用して、文献調査および探索的インタビューから作られた 19 の会議目標のセット [10] から最大 3 つを選択する。トピック分割については、評価用 LLM と同じ LLM を用いて行う。評価に十分なコンテキストを提供するために、スライディングウィンドウ方式を採用し、ターゲットセグメントとともに一定数の先行および後続セグメントを提示する。

5.2 セグメンテーションアライメント

セグメントレベルの有効性スコアをメタ評価する際の重要な課題は、モデルが予測した割境界と正解の分割境界の間の不一致である。セグメントの数と長さが異なることが多いため、直接的な一対一の比較は不可能である。これに対処するために、予測スコアを正解セグメントにマッピングするセグメンテーションアライメントを行う。

形式的には、正解セグメントを $\{[t_i, t_{i+1}]\}_{i=0}^{n-1}$、対応する有効性スコアを $\{e_{t_i}^{t_{i+1}}\}_{i=0}^{n-1}$ とする。同様に、予測セグメントを $\{[\hat{t}_j, \hat{t}_{j+1}]\}_{j=0}^{\hat{n}-1}$、対応する予測スコアを $\{\hat{e}_{\hat{t}_j}^{\hat{t}_{j+1}}\}_{j=0}^{\hat{n}-1}$ とする。各 i 番目の正解セグメントについて、それと時間的に重複するすべての予測セグメントから導出される単一のアライメント済み予測スコア $\bar{e}_{t_i}^{t_{i+1}}$ を計算する。このアライメント済みスコアは、重複する予測セグメントのスコアの加重平

	Context Window Size = 1		Context Window Size = 3	
モデル	**Spearman (ρ)**	**Kendall-Tau (τ)**	**Spearman (ρ)**	**Kendall-Tau (τ)**
Qwen3-32B (non-reasoning)	0.6445	0.4803	0.5996	0.4423
+ From Segmentation	0.2256	0.1604	0.2320	0.1644
+ From Speech	0.2180	0.1547	0.2493	0.1766
GPT-4o	0.6341	0.4756	0.5618	0.4153
+ From Segmentation	0.2360	0.1664	0.2412	0.1706
+ From Speech	0.2006	0.1430	0.2003	0.1421

表 1: 各入力設定における有効性スコアのメタ評価結果

均であり、重みは重複期間によって決定される。

$$\hat{e}_{t_i}^{t_{i+1}} = \frac{\sum_{j=0}^{\hat{n}-1} \hat{e}_{\hat{t}_j}^{\hat{t}_{j+1}} \cdot \Delta_{i,j}}{\sum_{j=0}^{\hat{n}-1} \Delta_{i,j}} \tag{1}$$

ここで、$\Delta_{i,j}$ は i 番目の正解セグメントと j 番目の予測セグメントの間の時間的重複の期間を表す。

このプロセスでは、重複する各予測セグメントの最終的なアライメント済みスコアへの寄与は、重複期間に比例する。このプロセスにより、正解スコアと並列なアライメント済み予測スコアのリストが得られ、相関メトリクスの直接計算が可能になる。

6 実験

6.1 実験設定

実験には、Qwen3-32B および GPT-4o を含む主要なオープンソースおよびプロプライエタリな LLM を選定し、推奨されるハイパーパラメータ設定を使用した。予測された有効性スコアと正解の有効性スコアとの間の相関係数を、セグメントレベルの Spearman (ρ) および Kendall-Tau (τ) を用いて評価した。ここで、各セグメントの正解スコアは、その 3 つのアノテーションスコアの平均とした。

6.2 実験結果

正解入力、モデル予測によるセグメント入力、および生の音声入力の 3 つの設定の下でフレームワークの性能を調査した。正解入力を用いた設定では、正解トピック分割と会議目標を提供し、上流タスクの性能の影響を排除した。また、コンテキストウィンドウサイズを 1 および 3 として調査を行った。結果を表 1 に示す。

正解分割を用いた場合、モデルは高い一貫性を示した。しかし、コンテキストについては、両モデルともウィンドウサイズ 1 が最適であった。これは、正解セグメントの完全性により、追加のコンテキストが不要であるためと推測される。ほかのモデルを含む比較結果を付録 B に示す。

一方、モデル予測による分割（+ From Segmentation）を用いた場合、正解分割を用いた場合と比較して相関係数が大幅に低下した。セグメントの不一致による性能低下を分離して評価するために、予測されたセグメントが与えられた場合の近似的な相関係数上限を計算した。これは、正解の有効性スコアを予測されたセグメントにアライメントし、これらのアライメントされたスコアを新しい予測として扱い、相関計算のために再度正解セグメントにアライメントし直すことで行った。その結果、Qwen3 のスピアマン上限は 0.6417、GPT-4o は 0.6828 であり、それぞれの Kendall-Tau の上限は 0.5095 と 0.5513 であった。これは、トピック分割の品質が最終的な相関に影響を与える重要な要因であることを示している。各モデルの分割性能については付録 C に示す。

最後に、ダイアライゼーションに pyannote [16]、ASR に Whisper-Large-V3 [17] を使用したエンドツーエンド（+ From Speech）は、正解トランスクリプトから開始した場合（+ From Segmentation）と同等の相関係数を達成した。これは、有効性評価コンポーネントが音声認識の誤りに対して比較的頑健であることを示唆しており、ノイズの多いテキストに対する LLM 固有の耐性によるものと考えられる。

7 おわりに

本研究では、会議有効性評価の新しいパラダイムを提案した。このパラダイムを実現するために、新しいメタ評価データセットを構築し、自動評価のための LLM ベースのフレームワークを提案した。実験ではベンチマークを確立し、LLM が人間の判断と強い相関を達成できることを実証した。エンドツーエンドシステムのベースラインを確立することで、本研究は会議分析および知的会議支援アシスタントにおける将来の研究のための強固な基盤を提供する。

謝辞

本研究は JST 次世代研究者挑戦的研究プログラム JPMJSP2110 および JSPS 科研費 JP23K28144 の助成を受けたものです.

参考文献

[1] Peter R Monge, Charles McSween, and JoAnne Wyer. **A profile of meetings in corporate America: Results of the 3M meeting effectiveness study**. Annenberg School of Communications, University of Southern California, 1989.

[2] Peter M Tobia and Martin C Becker. Making the most of meeting time. **Training and Development Journal**, Vol. 44, No. 8, pp. 34–38, 1990.

[3] N.C. Romano and J.F. Nunamaker. Meeting analysis: findings from research and practice. In **Proceedings of the 34th Annual Hawaii International Conference on System Sciences**, pp. 13 pp.–, 2001.

[4] Steven Rogelberg, Desmond Leach, Peter Warr, and Jennifer Burnfield. "not another meeting!" are meeting time demands related to employee well-being? **The Journal of applied psychology**, Vol. 91, pp. 83–96, 01 2006.

[5] Joseph A. Allen, NaleLehmann-Willenbrock. The key features of workplace meetings: Conceptualizing the why, how, and what of meetings at work. **Organizational Psychology Review**, Vol. 13, pp. 355 – 378, 2022.

[6] Yasaman Hosseinkashi, Lev Tankelevitch, Jamie Pool, Ross Cutler, and Chinmaya Madan. Meeting effectiveness and inclusiveness: Large-scale measurement, identification of key features, and prediction in real-world remote meetings. **Proc. ACM Hum.-Comput. Interact.**, Vol. 8, No. CSCW1, April 2024.

[7] Carol T Nixon and Glenn E Littlepage. Impact of meeting procedures on meeting effectiveness. **Journal of Business and Psychology**, Vol. 6, No. 3, pp. 361–369, 1992.

[8] Boni García, Micael Gallego, Francisco Gortázar, and Antonia Bertolino. Understanding and estimating quality of experience in webrtc applications. **Computing**, Vol. 101, No. 11, p. 1585–1607, November 2019.

[9] Ross Cutler, Yasaman Hosseinkashi, Jamie Pool, Senja Filipi, Robert Aichner, Yuan Tu, and Johannes Gehrke. Meeting effectiveness and inclusiveness in remote collaboration. **Proc. ACM Hum.-Comput. Interact.**, Vol. 5, No. CSCW1, April 2021.

[10] Willem Standaert, Steve Muylle, and Amit Basu. An empirical study of the effectiveness of telepresence as a business meeting mode. **Inf. Technol. and Management**, Vol. 17, No. 4, p. 323–339, December 2016.

[11] Marios Constantinides, Sanja Šćepanović, Daniele Quercia, Hongwei Li, Ugo Sassi, and Michael Eggleston. Comfeel: Productivity is a matter of the senses too. **Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.**, Vol. 4, No. 4, December 2020.

[12] Desmond Leach, Steven Rogelberg, Peter Warr, and Jennifer Burnfield. Perceived meeting effectiveness: The role of design characteristics. **Journal of Business and Psychology**, Vol. 24, pp. 65–76, 03 2009.

[13] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: a pre-announcement. In **Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction**, MLMI'05, p. 28–39, Berlin, Heidelberg, 2005. Springer-Verlag.

[14] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. **Journal of chiropractic medicine**, Vol. 15 2, pp. 155–63, 2016.

[15] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[16] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In **Interspeech 2023**, pp. 1983–1987, 2023.

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In **International conference on machine learning**, pp. 28492–28518. PMLR, 2023.

[18] Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. Unsupervised topic segmentation of meetings with bert embeddings, 2021.

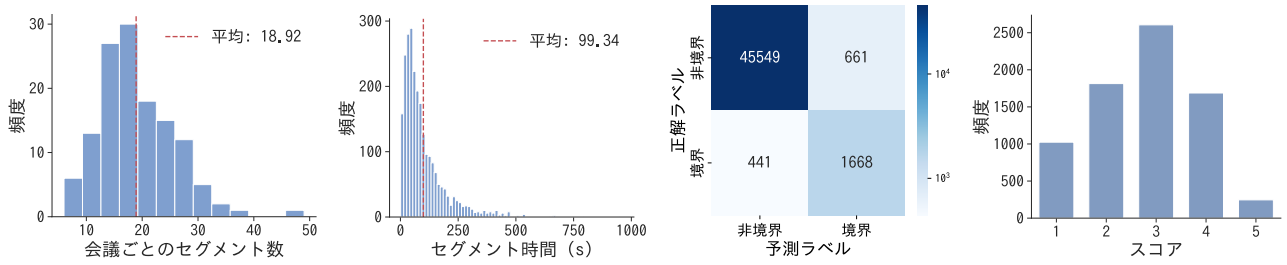
[19] Aleksei Artemiev, Daniil Parinov, Alexey Grishanov, Ivan Borisov, Alexey Vasilev, Daniil Muravetskii, Aleksey Rezvykh, Aleksei Goncharov, and Andrey Savchenko. Leveraging summarization for unsupervised dialogue topic segmentation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 4697–4704, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[20] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, pp. 2623–2631, 2019.

[21] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. **Mach. Learn.**, Vol. 34, No. 1–3, p. 177–210, February 1999.

[22] Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. **Computational Linguistics**, Vol. 28, No. 1, pp. 19–36, 03 2002.

A　AMI-ME データセットの統計



(a)　(b)　(c)　(d)

図 3: AMI-ME データセットの統計 (a) 会議ごとのセグメント数の分布 (b) セグメント時間の分布 (c) 隣接する発話間の境界特定を 2 値分類タスクとみなした場合のトピック分割の混同行列 (d) セグメント有効性スコアの分布

B　正解入力を用いた異なる LLM の評価

	Context Window Size = 1		Context Window Size = 3	
モデル	**Spearman (ρ)**	**Kendall-Tau (τ)**	**Spearman (ρ)**	**Kendall-Tau (τ)**
Llama3.3-70B-Instruct	0.6072	**0.4854**	0.5781	0.4633
DeepSeek-R1-70B	0.6132	0.4663	0.5820	0.4409
Qwen3-32B (reasoning)	0.6113	0.4578	0.5831	0.4356
Qwen3-32B (non-reasoning)	**0.6445**	0.4803	0.5996	0.4423
GPT-4o	0.6341	0.4756	0.5618	0.4153
Gemini-2.5-Flash (non-reasoning)	0.5624	0.4122	0.5260	0.3855

表 2: 正解入力設定における異なる LLM の有効性スコアのメタ評価結果

異なる LLM の有効性評価能力を分離しベンチマークするために、理想的な条件下で実験を行った。上流タスクの性能の影響を排除するため、全てのモデルに同一の正解トピック分割と会議目標を与えた。さらに、コンテキストウィンドウサイズ 1 および 3 について調査した。表 2 に示すように、結果は多くのモデルで高い一貫性を示しており、Qwen3-32B (non-reasoning) が最も高い Spearman 相関を達成した。Gemini-2.5-Flash は顕著な例外であり、著しく低い相関を示した。これは、最低スコアを過剰に使用する傾向があるためと考えられる。コンテキストウィンドウに関しては、すべてのモデルにおいて、サイズ 1 がサイズ 3 よりも一貫して良好な結果をもたらした。

C　トピック分割の実験

各トピック分割手法の性能を調査した。まず、境界なしのベースライン（「Absence」）、BERTSeg [18]、SumSeg [19] を含む様々な教師なしトピック分割モデルと、様々な LLM をベンチマークした。BERTSeg と SumSeg については、Optuna [20] を用いて公式実装でハイパーパラメータ探索を行った。トピック分割の評価指標として、Pk [21] および WindowDiff (Wd) [22] を採用した。表 3 に示すように、LLM は非 LLM ベースラインを大幅に上回る性能を示した。Gemini-2.5-Flash の性能が低いのは、過分割の傾向があるためであり、正解の 18.9 に対して会議あたり平均 28.7 セグメントを生成した。この結果によって、各 LLM の有効性評価は、その LLM 自身の分割出力に基づいて実施された。

モデル	**Wd ↓**	**Pk ↓**
Absence	0.4176	0.4176
BERTSeg	0.4139	0.4058
SumSeg	0.4069	0.4002
Llama3.3-70B-Instruct	0.4108	0.3727
Qwen3-32B (non-reasoning)	0.3795	0.3402
Qwen3-32B (reasoning)	0.3713	0.3418
DeepSeek-R1-70B	**0.3658**	**0.3325**
GPT-4o	0.3980	0.3548
Gemini-2.5-Flash (non-reasoning)	0.4486	0.3576

表 3: トピック分割の実験結果