CINESUBBENCH: EVALUATING LLMS ON LONG-FORM NARRATIVE AND CULTURAL UNDERSTANDING FROM MULTILINGUAL MOVIE SUBTITLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly evaluated in specialized domains such as law, medicine, software engineering, and cybersecurity, yet film remains comparatively underexplored despite requiring long-form narrative integration, multilingual interpretation, and culturally situated audience judgments. We introduce CineSubBench, a benchmark for evaluating long-context film understanding from multilingual movie subtitles. A subtitle track represents a film as thousands of short, temporally ordered utterances from which models must reconstruct characters, relationships, events, causal progression, and themes without explicit scene or event structure. CineSubBench contains 1,012 films with complete subtitle coverage in six languages, yielding 6,072 tracks and 8.13M timestamped subtitle entries. It provides a matched **multi-task, multilingual, and multicultural (MultiX)** evaluation setting: seven tasks span narrative reconstruction and abstraction, genre prediction, age suitability, country-specific motion-picture ratings across ten national classification systems, and subtitle-grounded language safety. Across nine LLMs, plot premises are recovered more reliably than event-complete synopses; cross-lingual consistency varies substantially across models and languages; national rating systems expose distinct calibration patterns; and strong profanity is far easier to ground than mild obscenity. CineSubBench establishes film as a long-context LLM evaluation domain and provides a unified benchmark for measuring narrative, multilingual, cultural, and evidence-grounding capabilities.

1 INTRODUCTION

Large language models are increasingly evaluated in specialized domains where general capabilities must transfer to domain-specific forms of evidence, reasoning, and decision making. Existing benchmarks span legal and clinical judgment (Guha et al., 2023; Singhal et al., 2023), finance and economics (Cao et al., 2024; Guo & Yang, 2024), software engineering (Jimenez et al., 2024), scientific and structured-data reasoning (Zhang et al., 2024; Duan et al., 2026), education and journalism (Hou et al., 2024; Li et al., 2024), and cybersecurity and geospatial reasoning (Dihan et al., 2025; Wang et al., 2026b). Film remains comparatively underexplored as a domain for LLM evaluation, despite combining several demanding capabilities: long-form narrative understanding, social interaction, implicit meaning, multilingual interpretation, and culturally situated audience judgments.

Film also presents a long-context language problem. Much existing long-context evaluation centers on articles, papers, books, retrieval collections, or other documents in which discourse or narrative structure is largely encoded in the text (Kryscinski et al., 2022; Liu et al., 2024; Goldman et al., 2024; Kim et al., 2024; Hamilton et al., 2026). A film unfolds differently. Its story develops over hours through dialogue, interactions among characters, and temporally separated events. Relationships evolve, motivations may remain implicit, and the significance of an early exchange can become clear only much later. Film understanding therefore requires a model to integrate evidence distributed throughout a long context and reconstruct the structure of an unfolding narrative.

Movie subtitles provide a distinctive representation of long-context film understanding. A subtitle track is a film-length sequence of short, timestamped utterances that preserves the dialogue and temporal progression through which the narrative develops. Dialogue carries character goals, re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

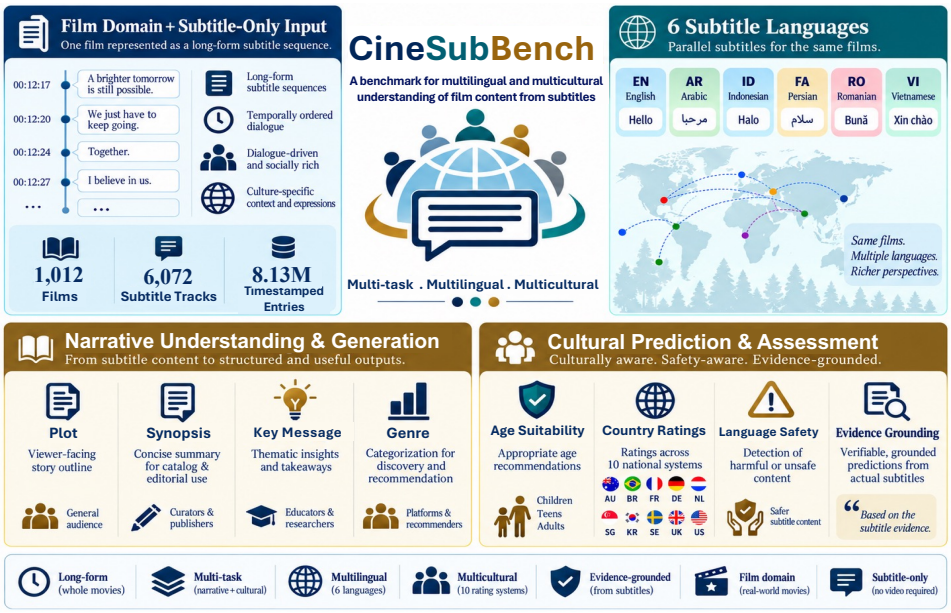


Figure 1: Overview of CineSubBench, a long-form MultiX benchmark for film understanding from subtitles. Each of 1,012 films has subtitle tracks in English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, yielding 6,072 tracks and 8.13M timestamped subtitle entries. Seven tasks span Narrative Understanding and Generation and Cultural Prediction and Assessment, with ratings across ten national classification systems.

lationships, decisions, reactions, conflicts, and consequences, allowing central narrative arcs and themes to be recovered from evidence across a film. Yet subtitles do not identify speakers, scene boundaries, narrative events, importance, motivations, or causal relations explicitly. Models must therefore reconstruct who is involved, what happened, how events connect, and which developments matter. This differs from both conventional document summarization and screenplay summarization: screenplays can provide scene descriptions, speaker attribution, stage directions, and explicit accounts of actions (Chen et al., 2022; Saxena & Keller, 2024), while timeline summarization can organize events around explicit dates or event markers (Steen & Markert, 2019). Subtitle timestamps provide temporal order without identifying the narrative units themselves, making subtitle-based film understanding more than compression of an already structured narrative.

Film also creates a natural setting for studying whether understanding remains consistent across linguistic and cultural contexts. The same story can be encountered through different subtitle languages, while audience guidance for that film may vary across national classification systems. These dimensions intersect in practice: a viewer may encounter a film through translated subtitles, seek an account of its narrative and themes, and rely on audience guidance defined within a particular national setting. Evaluating tasks, languages, and cultural judgments over the same films therefore provides a coherent basis for examining these dimensions together.

A subtitle-centered setting also enables consistent film-length evaluation across a broad range of LLMs. Full-video understanding introduces additional choices and computation for frame or scene selection, video decoding, audio processing and alignment, temporal selection, and cross-modal inference over hour-scale inputs. Measured performance can consequently reflect both narrative reasoning and choices in the audiovisual processing pipeline. Subtitles provide the same temporally ordered text interface to closed and open-weight language models, including systems without full-length video input, allowing the evaluation to focus directly on long-context narrative integration.

We introduce CineSubBench, organized around two challenges: **long-form understanding** and **MultiX evaluation**. The benchmark contains 1,012 films, 6,072 subtitle tracks, and 8.13M timestamped subtitle entries, with complete coverage in six subtitle languages and ten national motion-picture rating systems. Its **multi-task** dimension spans plot generation, spoiler-aware synopsis generation, key-message generation, multi-label genre prediction, age-suitability prediction, country-specific motion-picture rating prediction, and evidence-grounded language-safety assessment. Its **multilingual** dimension presents the same films through six subtitle languages, while its **multicultural** dimension retains ten national classification systems as distinct label spaces (Figure 1). Cross-

source entity resolution, complete-coverage optimization, subtitle verification, and task-specific curation support matched and auditable comparisons across these dimensions (Figure 2).

Across nine LLMs, CineSubBench reveals capability differences that aggregate rankings obscure. Models recover broad plot premises more reliably than event-complete synopses, cross-lingual consistency varies across models and languages, national rating systems expose distinct calibration behavior, and evidence-grounded language safety ranges from near-ceiling grounding of strong profanity to much weaker performance on mild obscenity. These results separate long-form narrative integration, multilingual consistency, culturally situated prediction, and evidence grounding as distinct dimensions of current LLM capability. We make three core contributions:

1. **A long-form task setting for LLM film understanding.** We formulate subtitle-based film understanding as a long-context language problem in which models reconstruct narrative structure from temporally ordered, dialogue-centered evidence distributed across an entire film. This setting enables evaluation of event selection, chronology, character tracking, causal integration, and thematic abstraction from a common film-length linguistic representation.
2. **A matched MultiX benchmark with broad coverage.** CineSubBench contains 1,012 films, 6,072 subtitle tracks, and 8.13M timestamped subtitle entries, with complete six-language subtitle coverage and ratings from ten national classification systems for every film. Seven tasks support **multi-task, multilingual, and multicultural** evaluation over the same film instances.
3. **A broad and deep evaluation of nine LLMs.** We evaluate nine models using task-specific metrics and matched comparisons across narrative tasks, subtitle languages, national rating systems, and subtitle-grounded safety evidence. The evaluation further examines task-level failure patterns, paired cross-lingual uncertainty, country-specific calibration, evidence-grounding errors, and qualitative cases that expose behaviors hidden by aggregate scores.

2 Related Work

Domain-grounded LLM evaluation. LLM benchmarks increasingly evaluate whether general-purpose models transfer to domains with specialized evidence and decision criteria, including law and medicine (Guha et al., 2023; Singhal et al., 2023), finance (Cao et al., 2024; Guo & Yang, 2024), software engineering (Jimenez et al., 2024), science and structured data (Duan et al., 2026; Zhang et al., 2024), education and journalism (Hou et al., 2024; Li et al., 2024), cybersecurity and geospatial reasoning (Dihan et al., 2025; Wang et al., 2026b), and multilingual web understanding (Awal et al., 2025). Film has received far less attention as a comparable LLM evaluation domain despite combining long-context narrative reasoning, social interaction, multilingual interpretation, and culturally situated judgments. CineSubBench addresses this gap with a unified film benchmark that evaluates these capabilities through its multi-task, multilingual, and multicultural design.

Long-context narrative and temporal understanding. Long input windows do not guarantee effective use of evidence distributed throughout a context (Liu et al., 2024; Goldman et al., 2024). Existing long-form benchmarks study summarization, faithfulness, narrative structure, and multilingual consistency in books and stories (Kryscinski et al., 2022; Kim et al., 2024; Hamilton et al., 2026; Kim et al., 2025), while culturally specific long-context evaluation shows that English-centered results need not transfer uniformly (Arora et al., 2025). Timeline summarization assumes explicit temporal events or markers (Steen & Markert, 2019); subtitle timestamps provide order without identifying event boundaries, causal structure, or narrative importance. CineSubBench instead evaluates film-length reconstruction from dialogue-centered evidence while holding the underlying films fixed across tasks, subtitle languages, and national classification systems.

Film, subtitle, and audiovisual understanding. Prior film and television work has addressed individual problems such as screenplay summarization (Chen et al., 2022; Saxena & Keller, 2024), narrative adaptation from movie dialogue (Shen & Garg, 2025), US age-rating (Shafaei et al., 2020), and content-severity prediction (Zhang et al., 2021). These settings do not provide a general LLM benchmark spanning long-form narrative generation, genre, age suitability, multiple national rating systems, multilingual subtitle inputs, and subtitle-indexed safety evidence over the same films. Audiovisual benchmarks such as MovieQA, LongVideoBench, InfiniBench, MF^2, and ShotBench instead target visual grounding, multimodal reasoning, event understanding, scene interpretation, or cinematographic analysis (Tapaswi et al., 2016; Wu et al., 2024; Ataallah et al., 2025; Zaranis et al.,

2025; Liu et al., 2025; Wang et al., 2026a). Copyright, redistribution, and audiovisual-processing requirements can also limit scale and access, with some benchmarks relying on selected clips, existing video assets, or separate media acquisition. CineSubBench instead uses publicly accessible subtitle tracks, enabling matched MultiX comparison across tasks, languages, national classification systems, and model families, providing a common benchmark for measuring progress.

3 CINESUBBENCH

CineSubBench is organized around two properties: **long-form input** and **MultiX evaluation**. It contains 1,012 films released between 1977 and 2026, represented by 6,072 subtitle tracks and 8.13M timestamped subtitle entries in English, Arabic, Indonesian, Persian, Romanian, and Vietnamese. Every film also has ratings from the same ten national motion-picture classification systems, enabling matched comparisons across tasks, languages, and countries.

Each subtitle input is film-length: mean prompt-formatted tracks contain roughly 40K–45K tokens across languages, with the longest exceeding 100K. The difficulty arises from both length and structure, since evidence appears as short, temporally ordered dialogue rather than continuous exposition. We use *long-form* for the input and *long-context understanding* for the capability required to integrate evidence distributed across it. Each subtitle entry retains its index, time interval, and text, supporting both film-level understanding and line-level grounding. The benchmark is **multi-task**, spanning seven narrative and cultural tasks; **multilingual**, with the same films presented in six subtitle languages; and **multicultural**, with audience classification evaluated under ten distinct national rating systems. Appendix A.9 reports corpus, input-length, label, and subtitle-duration statistics.

3.1 NARRATIVE UNDERSTANDING AND GENERATION

The narrative tasks probe levels of understanding from the same film-length subtitle sequence. We distinguish plot, synopsis, and key message by their function rather than by output length.

Plot Generation. The model generates a concise premise identifying the central character or group, setup, conflict, and driving situation, emphasizing the dominant narrative over retelling.

Synopsis Generation. The model generates a spoiler-aware account of the narrative arc, including developments, character goals and conflicts, turning points, climax, and resolution. Compared with plot generation, this requires event selection, chronology, causality, and character tracking.

Key-Message Generation. The model generates one concise moral, social, or emotional takeaway, requiring thematic abstraction beyond simply recounting narrative events.

Genre Prediction. The model predicts one or more genres from a normalized 22-label space using evidence from story, character, setting, tone, conflict, and broader thematic and stylistic patterns. Appendix B.3 provides the complete task instructions and response schemas.

3.2 CULTURAL PREDICTION AND ASSESSMENT

These tasks evaluate how models map film-level subtitle evidence to audience-oriented judgments defined by different national and institutional target settings and classification frameworks.

Age-Suitability Prediction. The model predicts the Common Sense Media minimum recommended age. This family-oriented suitability judgment reflects perceived developmental appropriateness and is evaluated separately from formal national motion-picture classification.

Country-Specific Motion-Picture Rating Prediction. For every film, CineSubBench includes ratings from Australia, Brazil, France, Germany, the Netherlands, Singapore, South Korea, Sweden, the United Kingdom, and the United States. Each country retains its own ordered label space because these classifications reflect distinct institutional standards rather than interchangeable age labels. The task therefore evaluates how the same film evidence maps into ten national classification systems. Appendix Table 4 lists the normalized label spaces.

Subtitle-Grounded Language-Safety Assessment. Language safety is evaluated from the English subtitle track in four lexical categories: strong profanity, crude bodily language, mild obscenity, and religious profanity/exclamation. Models predict category counts and exact subtitle indices

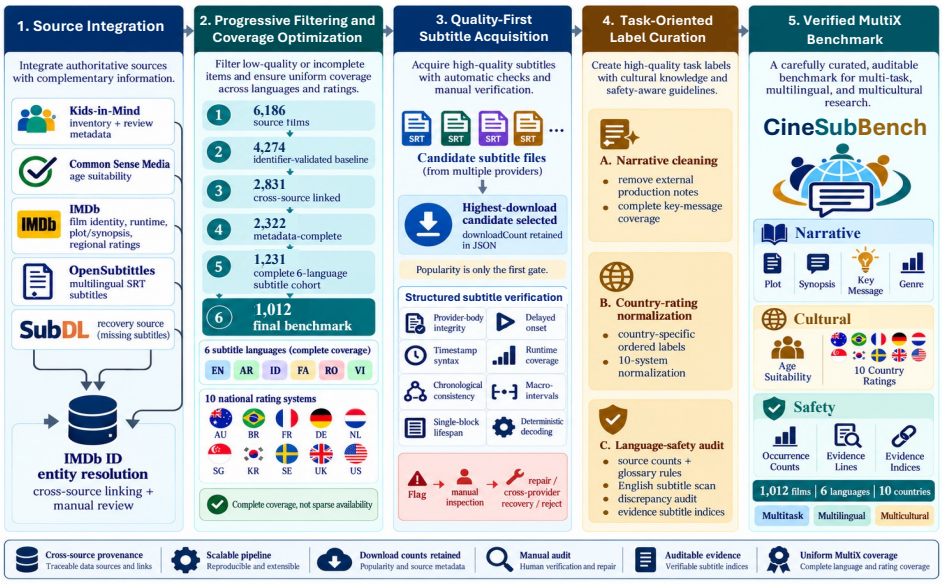


Figure 2: Construction and quality-assurance pipeline for CineSubBench. Cross-source entity resolution and filtering reduce the initial 6,186-film inventory to a metadata-complete cohort. Complete-coverage optimization then selects six subtitle languages and ten national rating systems. Subtitle candidates undergo structural, temporal, encoding, language-consistency, and runtime checks with manual inspection and cross-provider recovery where needed. Task-oriented curation yields the final matched MultiX benchmark of 1,012 films.

supporting them. Gold labels are constructed from audited subtitle occurrences, while Kids-in-Mind counts and glossary definitions are retained as provenance and an audit signal. Requiring evidence indices separates evidence recovery from category assignment and makes missed or unsupported judgments directly inspectable. Appendix A.6 details the annotation and audit procedure.

3.3 Construction and Quality Assurance

CineSubBench integrates film metadata, audience guidance, national classifications, and subtitle assets through IMDb title identifiers. Kids-in-Mind provides the initial film inventory and language-content metadata, Common Sense Media provides age-suitability labels and storylines, IMDb provides narrative metadata, runtimes, and regional ratings, and OpenSubtitles provides the subtitle source, with SubDL used for recovery. As Figure 2 shows, cross-source entity resolution and task-metadata filtering reduce the initial 6,186 films before complete-coverage optimization selects a six-language cohort of 1,231 films and a final 1,012-film cohort with ratings from ten national systems. Language and country sets are selected for shared coverage over the same films rather than marginal frequency alone. Appendix A.4 details the optimization procedure and coverage analysis.

For film-language pairs with multiple subtitle candidates, source-reported download count provides a selection signal, followed by checks for provider-error payloads, malformed SRT structure, temporal inconsistencies, delayed starts, large gaps, insufficient runtime coverage, encoding failures, and language inconsistencies. Automated flags trigger manual inspection, repair, or cross-provider recovery where needed. Task references undergo curation: narrative fields are cleaned to retain film-internal content, age suitability remains distinct from national classification, country ratings are normalized within their national systems, and language-safety labels are audited against indexed subtitle evidence. Appendices A.5 and A.7 document the verification and curation procedures.

4 Evaluation Setup

Models and input regimes. Our main evaluation compares nine models across three access groups: closed frontier models (GPT-5.6 Sol, Gemini 3.8 Flash, and Claude Haiku 4.5), closed baselines (GPT-5 Nano and Gemini 3.5 Flash Lite), and open-weight models (DeepSeek V4 Pro 1.6T, Qwen3 235B, Mistral 4 119B, and Llama 4 Scout 17B). For the six-language evaluation, the same films are presented through English, Arabic, Indonesian, Persian, Romanian, and Vietnamese

subtitles. Narrative and cultural predictions are evaluated separately for each language, while narrative generations are written in English. We use CL for the equal-weight mean of the five non-English settings; subtitle-grounded language-safety assessment is English-only because its gold evidence is localized to the English track. Appendix Tables 15 and 16 report model coverage and generation settings, and Appendix G presents the English-only within-family scaling analysis.

LLM-as-a-judge evaluation. Open-ended plot, synopsis, and key-message generations are evaluated with GPT-5.6 Luna as a common reference-based LLM judge (Liu et al., 2023; Gu et al., 2026; Li et al., 2025). Because narrative evaluation spans long film contexts across multiple models, tasks, subtitle languages, and thousands of generations, judge inference is a substantial component of evaluation cost. We therefore use Luna as a fast, cost-efficient judge while applying the same references, task-specific rubrics, and structured scoring protocol to every model. Appendix B.2 provides the scoring definitions, Appendix B.3 gives the full judge instructions, and Appendix B.4 reports agreement with an independent narrative judge, with 98.4% within-one-point agreement overall.

4.1 Evaluation Metrics

We use task-specific metrics. Plot, synopsis, and key-message generations are scored on a reference-based **1–5 rubric** with GPT-5.6 Luna; narrative overall is their equal-weight mean. Genre prediction uses micro-F1. Age suitability uses MAE in years together with exact and near-match accuracy. Country-specific ratings use exact match and ordinal MAE computed within each national label space. Appendix B.2 provides definitions and secondary measures, while Appendix B.3 gives the task and evaluator instructions. For language safety, category-agnostic evidence F1 measures recovery of relevant subtitle indices regardless of category, while strict F1 additionally requires the correct lexical category. In the leaderboard, LC denotes this English-only language-content assessment.

Cross-task composite. We use a composite only as a compact summary across task families; subsequent analyses retain the original task metrics. Let S_{nar} denote narrative overall, $F1_{\mathrm{genre}}$ genre micro-F1, $\mathrm{MAE}_{\mathrm{age}}$ age MAE, $\mathrm{EM}_{\mathrm{country}}$ country-rating exact match, and $F1_{\mathrm{LC}}$ strict language-safety evidence F1. Percentage metrics are expressed on a 0–100 scale:

$$\begin{aligned} N_{\mathrm{nar}} &= 25(S_{\mathrm{nar}} - 1), & N_{\mathrm{age}} &= 100\left(1 - \frac{\min(\mathrm{MAE}_{\mathrm{age}}, 16)}{16}\right), \\ N_{\mathrm{cult}} &= \frac{N_{\mathrm{age}} + \mathrm{EM}_{\mathrm{country}}}{2}, & S_{\mathrm{overall}} &= \frac{N_{\mathrm{nar}} + F1_{\mathrm{genre}} + N_{\mathrm{cult}} + F1_{\mathrm{LC}}}{4}. \end{aligned} \tag{1}$$

The value 16 is the span of the age-label space from 2+ to 18+. Narrative, genre, age, and country metrics are first averaged equally over the six subtitle languages, whereas $F1_{\mathrm{LC}}$ is English-only.

5 Results

We organize the evaluation around *four* questions: whether aggregate ranking reflects uniform capability across tasks (RQ1), how reliably understanding carries across subtitle languages (RQ2), where cultural calibration differs across national settings and classification systems (RQ3), and whether language-safety judgments can be grounded in auditable subtitle evidence (RQ4).

5.1 RQ1: Does a Single Rank Reflect Uniform Capability?

Aggregate ranking hides different capability profiles. Table 1 summarizes performance across the benchmark. Gemini 3.8 Flash reaches the highest composite score at 78.04, followed by GPT-5.6 Sol at 74.70 and DeepSeek V4 Pro at 65.52. Paired film-level comparisons of these leading composite scores are reported in Appendix C.2. Their task-level profiles, however, differ substantially. Sol achieves the strongest synopsis score (3.40), whereas Gemini leads on plot (4.16), key message (3.11), genre micro-F1 (84.93%), country-rating exact match (77.87%), and strict language-safety evidence F1 (77.8%). A similar aggregate position therefore need not reflect the same strengths in narrative reconstruction, classification, cultural calibration, and evidence grounding.

Strong performance transfers unevenly across tasks. DeepSeek V4 Pro exceeds Sol on genre micro-F1 (81.63% versus 81.55%) and age accuracy within one year (79.33% versus 78.75%), with

Table 1: All-language comparison across the nine models with complete benchmark coverage. Narrative, genre, age, and country metrics are averaged equally over English, Arabic, Indonesian, Persian, Romanian, and Vietnamese subtitle inputs; language-content (LC) evidence metrics are English-only. Models are ordered by the composite score in Equation 1. LC agnostic F1 matches safety-relevant subtitle evidence regardless of category, while strict F1 also requires the correct lexical category. Best values are bold and second-best distinct values are underlined; background shading is normalized within each metric column.

Rank	Model	Family	Composite	Narrative Understanding and Generation				Cultural Prediction and Assessment					LC Assessment	
			Overall score (0–100) ↑	Plot ↑	Synopsis ↑	Key message ↑	Overall ↑	Genre micro-F1 (%) ↑	Age exact (%) ↑	Age ±1 (%) ↑	Age MAE ↓	Country EM (%) ↑	Country MAE ↓	LC agnostic F1 (%) ↑	LC strict F1 (%) ↑
🥇	Gemini 3.8 Flash	Closed frontier	**78.04**	**4.16**	3.34	**3.11**	**3.54**	**84.93**	**34.25**	**81.42**	**0.93**	**77.87**	**0.31**	**86.8**	**77.8**
🥈	GPT-5.6 Sol	Closed frontier	74.7	4.1	**3.4**	3.08	3.52	81.55	31.08	78.75	0.98	65.42	0.44	81.7	74.5
🥉	DeepSeek V4 Pro 1.6T	Open-weight	65.52	3.91	2.93	2.9	3.25	81.63	30.67	79.33	0.98	58.03	0.52	57.5	48.3
4	Gemini 3.5 Flash Lite	Closed baseline	64.7	3.83	2.91	2.99	3.25	80.92	31.83	72.17	1.09	64.65	0.46	65.3	42.8
5	Claude Haiku 4.5	Closed frontier	57.5	3.45	2.44	2.84	2.91	71.45	27.42	72.75	1.16	54.32	0.57	58.1	37.2
6	Qwen3 235B	Open-weight	53.09	3.36	2.3	2.76	2.8	68.28	22.0	59.25	1.47	47.88	0.64	45.0	29.6
7	Mistral 4 119B	Open-weight	50.67	2.8	2.0	2.52	2.44	66.83	19.42	54.5	1.78	47.37	0.68	49.6	31.7
8	GPT-5 Nano	Closed baseline	46.67	2.97	1.9	2.64	2.5	72.43	20.67	60.75	1.56	43.87	0.76	21.7	9.6
9	Llama 4 Scout 17B	Open-weight	41.58	2.21	1.67	2.2	2.03	66.92	18.58	50.25	1.93	34.12	0.93	28.4	12.7

both at 0.98-year MAE. Larger gaps appear in narrative overall (3.25 versus 3.52), country-rating exact match (58.03% versus 65.42%), and language-safety evidence F1 (48.3% versus 74.5%). Its third-place composite therefore reflects a different capability balance rather than a uniform gap from the leading models. English-only and five-language-average results appear in Appendix C.1.

5.2 RQ2: How Robust Is Understanding Across Subtitle Languages?

Cross-lingual consistency varies substantially across models. Narrative performance decreases from English to the five-language mean for all nine models, but the magnitude ranges from −0.05 points for GPT-5.6 Sol to −0.28 for Llama 4 Scout on the 1–5 scale. Paired intervals for the smallest changes include zero, whereas the larger declines are more clearly separated (Appendix Table 37). The detailed English-to-cross-lingual comparison across narrative, genre, age, and country ratings appears in Appendix D.1.

Persian is the most consistent observed stress setting. Figure 3 shows that Persian yields the lowest narrative-overall score for every model, the lowest genre score for eight of nine, and the lowest country-rating exact match for every model. Sol drops from 3.56 in English to 3.41 in Persian on narrative overall, while Llama 4 Scout falls from 2.26 to 1.79; country exact match drops from 67.2% to 62.4% and 33.1% to 31.8%, respectively. With films and references held constant, these results reflect end-to-end consistency across subtitle languages.

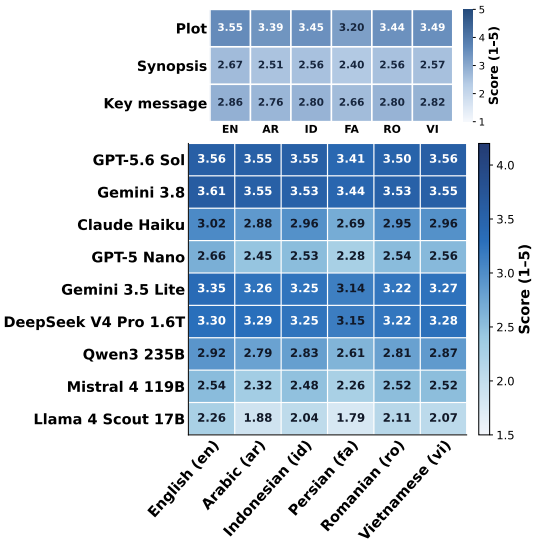


Figure 3: Narrative performance across subtitle languages. Plot remains more reliably recovered than synopsis, while Persian is the lowest observed narrative setting across the evaluated models.

Premise recovery is more robust than full narrative reconstruction. Across models, English averages are 3.55 for plot, 2.67 for synopsis, and 2.86 for key message; Persian averages are 3.20, 2.40, and 2.66. The gap persists for strong models: Sol scores 4.10 versus 3.40 and DeepSeek 3.91 versus 2.93 on plot and synopsis. Evaluator annotations show that 79.7% of synopses omit a major event and 71.1% invent one. Full results and error analyses appear in Appendix D.2.

5.3 RQ3: Where Does Cultural Calibration Fail?

Similar age accuracy can conceal opposite calibration tendencies. Across six subtitle languages, Gemini 3.8 Flash matches the exact reference age in 34.2% of cases and falls within one year in 81.4%; GPT-5.6 Sol reaches 31.1% and 78.8%, respectively. Nearly half of both models'

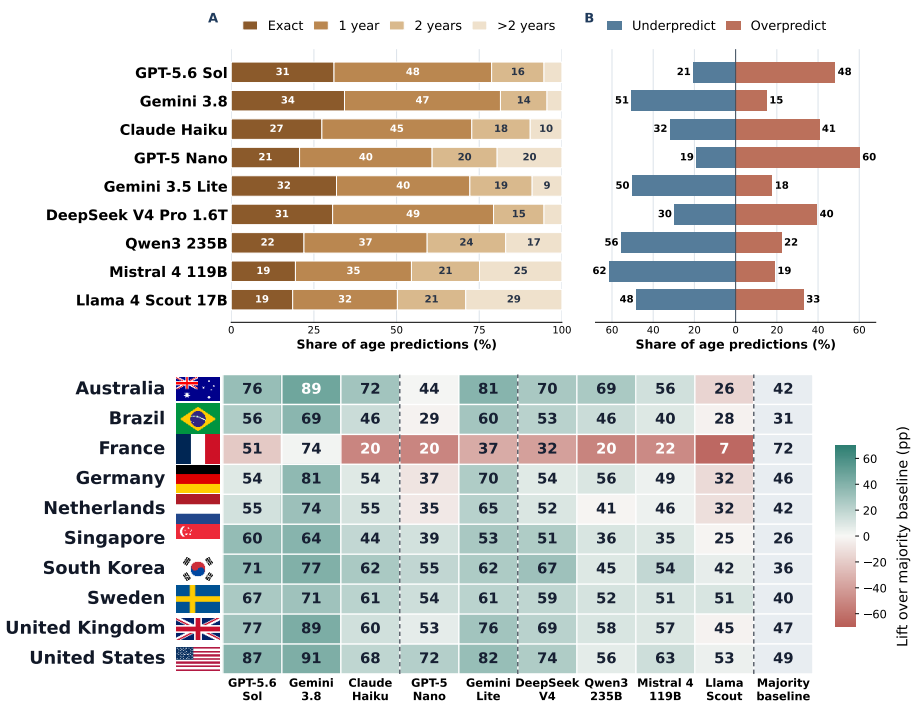


Figure 4: Cultural prediction reveals distinct calibration patterns. **Top:** Age errors by magnitude and direction. **Bottom:** Country-rating exact match across six subtitle languages, with shading showing lift over each country's majority baseline. France pairs a high baseline with weak performance for most models, while Brazil shows that lower raw accuracy can still yield substantial baseline gains.

predictions miss by exactly one year, but in opposite directions: Sol overpredicts more often than it underpredicts (48.3% versus 20.6%), whereas Gemini shows the reverse (15.2% versus 50.6%). Thus, similar near-match accuracy can reflect substantially different audience-guidance behavior. Appendix E.1 reports the full error-magnitude and direction analysis.

National rating systems reveal distinct calibration patterns. Figure 4 shows why country accuracy must be interpreted within each national label distribution. France has 31.6% mean model exact match against a 72.0% majority baseline; eight of nine models fall below it, and 95.3% of nonzero ordinal errors assign a stricter rating. Gemini reaches 73.8%, only slightly above baseline. Brazil differs: its largest class covers 31.0% of films, while Sol reaches 56.1%, a 25.1-point gain. Thus, similar raw accuracies can reflect very different calibration across national systems. Gemini and Sol achieve six-language country MAE of 0.31 and 0.44, respectively.

Cultural prediction also changes across subtitle languages. The direction is not uniform across models and countries. Romanian subtitles raise country-rating exact match by 3.7 points for Llama 4 Scout and 3.4 for GPT-5 Nano. Averaged across models, Romanian improves Brazil and Germany by 2.3 and 1.7 points, respectively, while all five non-English settings reduce US performance. Complete model-language and country-language analyses appear in Appendix E.2.

5.4 RQ4: Can Safety Judgments Be Evidence-Grounded?

Evidence recovery and category assignment are distinct capabilities. Figure 5 separates category-agnostic evidence recovery from strict matching, which also requires the correct lexical category. Gemini reaches 86.8% agnostic F1 and 77.8% strict F1, while Sol reaches 81.7% and 74.5%. This gap shows that models can locate relevant dialogue yet assign it to the wrong category. Precision and recall reveal different behaviors: Sol recovers 96.3% of gold evidence at 70.9% precision, whereas DeepSeek reaches 77.8% precision but only 45.6% recall.

Grounding difficulty differs sharply by lexical category. Strong profanity reaches 99.6% strict F1 for Sol and 99.1% for Gemini, compared with 39.7% and 52.0% for mild obscenity. DeepSeek

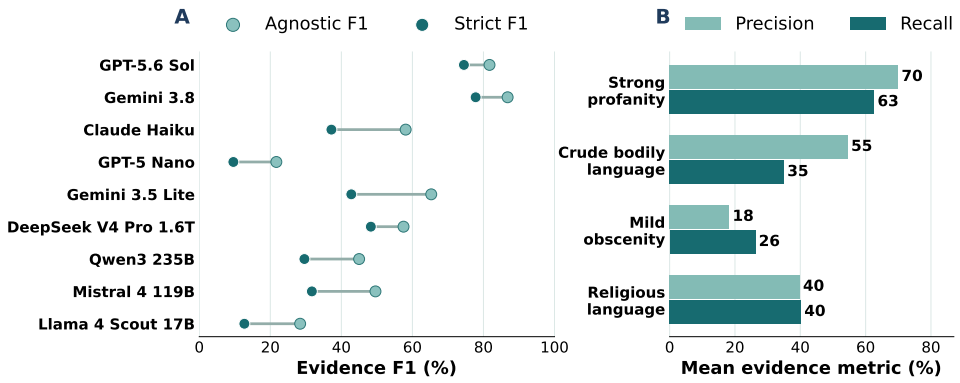


Figure 5: Evidence-grounded language-safety performance. **(A)** Category-agnostic F1 measures recovery of relevant subtitle evidence; strict F1 also requires the correct lexical category. **(B)** Mean precision and recall by category across models. Strong profanity is substantially easier to ground than mild obscenity.

shows the same ordering at 68.2% versus 22.6%. Across nine models, strong profanity reaches 62.8% category-level F1 versus 20.3% for mild obscenity, with 73.7% of its gold evidence missed. Precision, recall, count-error, category-specific, and false-positive analyses appear in Appendix F.1.

6 Discussion

Long-form film understanding remains challenging. Even frontier models do not uniformly recover a film from its subtitle sequence. Premises are easier than complete narrative arcs: synopsis scores remain around 3.4/5 for strong models, with major-event omissions and invented events. This leaves substantial headroom for applications such as film summarization, search, catalog enrichment, and narrative analysis, which require tracking characters, events, and causal progression.

Multilingual access does not ensure multilingual reliability. Changing only the subtitle language produces shifts in model behavior. Narrative performance declines from English to the non-English settings for all nine models in point estimate, while Persian repeatedly yields the weakest results across several tasks. CineSubBench therefore exposes cross-lingual instability that English-only film evaluation would miss, a limitation for systems serving multilingual audiences.

Narrative competence does not guarantee cultural calibration. Strong narrative performance can coexist with weak audience-guidance prediction. In France, mean country-rating exact match is only 31.6% against a 72.0% majority baseline, whereas Brazil shows substantial gains over its weaker baseline. Evaluating the same films across ten national systems reveals meaningful country-specific calibration differences that a single generic rating task would obscure.

Evidence grounding reveals further headroom. Film-facing safety assessment requires both a judgment and evidence supporting it. While strong profanity approaches near-ceiling grounding for the best models, mild obscenity reaches only 52.0% strict F1 for Gemini and 39.7% for Sol. By jointly evaluating narrative understanding, multilingual consistency, cultural calibration, and evidence grounding, CineSubBench shows that current LLMs remain far from uniformly reliable for film understanding and provides a common benchmark for measuring progress.

7 Conclusion

We introduced CineSubBench, a benchmark for long-form film understanding under a multi-task, multilingual, and multicultural evaluation setting. Across nine LLMs, premise recovery is more reliable than complete narrative reconstruction, cross-lingual consistency varies across models and languages, national classification systems reveal distinct calibration behavior, and evidence-grounded language safety remains strongly category-dependent. These findings show substantial headroom for current models across narrative, linguistic, cultural, and evidence-grounding dimensions of film understanding. By combining film-length subtitles with matched task, language, and cultural coverage, CineSubBench provides a unified benchmark for measuring progress across these capabilities.

AI USE STATEMENT

We used generative AI tools in limited assistive roles. Specifically, they were used for **(1) text editing**, including improvements to grammar, spelling, phrasing, and word choice; **(2) code assistance**, including code fragments for data preprocessing and filtering, routine implementation tasks, and debugging; and **(3) visual assistance**, limited to generating decorative icons used in figs. 1 and 2.

All AI-assisted text, code, and visual materials were reviewed, adapted, and approved by the authors, and any generated or modified code was manually verified and tested before use. The authors developed the research questions and methodology, designed the experiments, conducted the analyses, interpreted the results, and determined the final presentation and conclusions. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

Research scope and public data. CineSubBench is intended for non-commercial research and benchmarking. Its construction uses publicly accessible film metadata, audience-guidance information, national classification labels, and community-contributed subtitle files. Source roles and provenance are documented in Appendix A.1, with subtitle acquisition and recovery detailed in Appendix A.3. The benchmark uses these resources to study model capabilities over film-length linguistic evidence and does not distribute the underlying audiovisual works.

Subtitle-based film evaluation. Film understanding poses a distinctive data-access challenge because the underlying audiovisual works are copyrighted and costly to store, process, and redistribute at scale. Existing audiovisual benchmarks such as MovieQA, LongVideoBench, InfiniBench, MF^2, and ShotBench commonly operate over selected films, clips, frames, or externally obtained video assets (Tapaswi et al., 2016; Wu et al., 2024; Ataallah et al., 2025; Zaranis et al., 2025; Liu et al., 2025; Wang et al., 2026a). CineSubBench instead builds on publicly accessible, community-contributed subtitle tracks, enabling film-length evaluation over 1,012 films without redistributing movie video. This representation also makes matched evaluation across six languages, ten national classification systems, and diverse LLM families practical while retaining temporally ordered narrative evidence.

Privacy and data sanitization. We apply an automated, auditable sanitization procedure to all six subtitle tracks as part of the quality-assurance pipeline (Appendix A.5). Across 8,133,088 subtitle entries, 6,594 non-narrative distribution artifacts are removed, including subtitle-team credits, contact details, social-media handles, advertisements, download promotions, and external promotional links. In 65 otherwise narrative-relevant entries containing an email address, the surrounding text is retained and only the address is replaced with `[email redacted]`. Film identifiers, timestamps, language metadata, and original subtitle indices remain unchanged.

Content and intended use. Films naturally contain profanity, mature themes, and other sensitive language relevant to the benchmark tasks. Language-safety annotations preserve this task-relevant evidence while localizing it through subtitle indices rather than reproducing offensive expressions in the label fields; Appendix A.6 documents their construction and audit. Field-level provenance and the released data structure are described in Appendix A.8. CineSubBench is intended for research on long-context, multilingual, and culturally situated film understanding.

REPRODUCIBILITY STATEMENT

The paper and Appendix document dataset construction and quality assurance, model and task coverage, generation settings, evaluation metrics, prompts, and robustness analyses. All benchmark data and experimental code are available at **`anonymous.4open.science/r/CineSubBench`**, including data-processing, model-inference, evaluation, scoring, and analysis code. The current release is nearly complete for review; a polished and fully documented version will be released upon the paper's acceptance.

REFERENCES

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11772–11817, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.578. URL https://aclanthology.org/2025.acl-long.578/.

Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. InfiniBench: A benchmark for large multi-modal models in long-form movies and TV shows. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19485–19512, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.984. URL https://aclanthology.org/2025.emnlp-main.984/.

Rabiul Awal, Mahsa Massoud, Aarash Feizi, Zichao Li, Suyuchen Wang, Christopher Pal, Aishwarya Agrawal, David Vazquez, Siva Reddy, Juan A. Rodriguez, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. WebMMU: A benchmark for multimodal multilingual website understanding and code generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25118–25145, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1276. URL https://aclanthology.org/2025.emnlp-main.1276/.

Tianyu Cao, Natraj Raman, Danial Dervovic, and Chenhao Tan. Characterizing multimodal long-form summarization: A case study on financial reports. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=hDoN0CAy5e.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. SummScreen: A dataset for abstractive screenplay summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8602–8615, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.589. URL https://aclanthology.org/2022.acl-long.589/.

Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. MapEval: A map-based evaluation of geo-spatial reasoning in foundation models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13774–13813. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/dihan25a.html.

Haonan Duan, Stephen Zhewen Lu, Caitlin Fiona Harrigan, Nishkrit Desai, Jiarui Lu, Michał Koziarski, Leonardo Cotta, and Chris J. Maddison. Measuring scientific capabilities of language models with a systems biology dry lab. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=Cmx6b7w2nk.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. Is it really long context if all you need is retrieval? towards genuinely difficult long context NLP. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16576–16586, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.924. URL https://aclanthology.org/2024.emnlp-main.924/.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *The Innovation*, 7(6): 101253, 2026. ISSN 2666-6758. doi: https://doi.org/10.1016/j.xinn.2025.101253. URL https://www.sciencedirect.com/science/article/pii/S2666675825004564.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44123–44279. Curran Associates, Inc., 2023. doi: 10.52202/075280-1915. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/89e44582fd28ddfea1ea4dcb0ebbf4b0-Paper-Datasets_and_Benchmarks.pdf.

Yue Guo and Yi Yang. EconNLI: Evaluating large language models on economics reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 982–994, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.58. URL https://aclanthology.org/2024.findings-acl.58/.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. NarraBench: A comprehensive framework for narrative benchmarking. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3786–3801, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.176. URL https://aclanthology.org/2026.eacl-long.176/.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. E-EVAL: A comprehensive Chinese k-12 education evaluation benchmark for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7753–7774, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.462. URL https://aclanthology.org/2024.findings-acl.462/.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=YfHxQSoaWU.

Yekyung Kim, Jenna Russell, Marzena Karpinska, and Mohit Iyyer. One ruler to measure them all: Benchmarking multilingual long-context language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3vxxB3Ar9r.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. BOOKSUM: A collection of datasets for long-form narrative summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6536–6558, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.488. URL https://aclanthology.org/2022.findings-emnlp.488/.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

Miao Li, Ming-Bin Chen, Bo Tang, Shengbin Hou, Pengyu Wang, Haiying Deng, Zhiyu Li, Feiyu Xiong, Keming Mao, Peng Cheng, and Yi Luo. NewsBench: A systematic evaluation framework for assessing editorial capabilities of large language models in Chinese journalism. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9993–10014, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.538. URL https://aclanthology.org/2024.acl-long.538/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Hongbo Liu, Jingwen He, Yi Jinn, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Weichao Chen, Yu Qiao, Wanli Ouyang, Shengjie Zhao, and Ziwei Liu. Shotbench: Expert-level cinematic understanding in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=1DgSkx8L63.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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 (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Rohit Saxena and Frank Keller. MovieSum: An abstractive summarization dataset for movie screenplays. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4043–4050, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.239. URL https://aclanthology.org/2024.findings-acl.239/.

Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar, and Thamar Solorio. Age suitability rating: Predicting the MPAA rating based on movie dialogues. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1327–1335, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.166/.

Siqi Shen and Amanmeet Garg. Adapting large language models for movie domain with narrative understanding tasks. In Gemma Boleda and Michael Roth (eds.), *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 187–200, Vienna, Austria, July 2025.

Association for Computational Linguistics. ISBN 979-8-89176-271-8. doi: 10.18653/v1/2025.conll-1.13. URL https://aclanthology.org/2025.conll-1.13/.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Julius Steen and Katja Markert. Abstractive timeline summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 21–31, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5403. URL https://aclanthology.org/D19-5403/.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4631–4640, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Tapaswi_MovieQA_Understanding_Stories_CVPR_2016_paper.html.

Xinran Wang, Songyu Xu, Shan Xiangxuan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Yanhua huang, Kongming Liang, and Zhanyu Ma. Cinetechbench: A benchmark for cinematographic technique understanding and generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026a. URL https://openreview.net/forum?id=LBKyjz2ESc.

Zhun Wang, Tianneng Shi, Jingxuan He, Matthew Cai, Jialin Zhang, and Dawn Song. Cybergym: Evaluating AI agents' real-world cybersecurity capabilities at scale. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=2YvbLQEdYt.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/329ad516cf7a6ac306f29882e9c77558-Paper-Datasets_and_Benchmarks_Track.pdf.

Emmanouil Zaranis, António Farinhas, Saul Santos, Beatriz Canaverde, Miguel Moura Ramos, Aditya K Surikuchi, André Viveiros, Baohao Liao, Elena Bueno-Benito, Nithin Sivakumaran, Pavlo Vasylenko, Shoubin Yu, Sonal Sannigrahi, Wafaa Mohammed, Ben Peters, Danae Sánchez Villegas, Elias Stengel-Eskin, Giuseppe Attanasio, Jaehong Yoon, Stella Frank, Alessandro Suglia, Chrysoula Zerva, Desmond Elliott, Mariella Dimiccoli, Mohit Bansal, Oswald Lanz, Raffaella Bernardi, Raquel Fernández, Sandro Pezzelle, Vlad Niculae, and André F. T. Martins. Movie facts and fibs (mf^2): A benchmark for long movie understanding, 2025. URL https://arxiv.org/abs/2506.06275.

Shiyue Zhang and Mohit Bansal. Finding a balanced degree of automation for summary evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6617–6632, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.531. URL https://aclanthology.org/2021.emnlp-main.531/.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. TableLlama: Towards open large generalist models for tables. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6024–6044, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.335. URL https://aclanthology.org/2024.naacl-long.335/.

Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and Thamar Solorio. From none to severe: Predicting severity in movie scripts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3951–3956, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.332. URL https://aclanthology.org/2021.findings-emnlp.332/.

Supplementary Material: Appendices

Appendix Contents

A Dataset Construction and Quality Assurance
- A.1 Source roles and construction overview
- A.2 Entity resolution and progressive filtering
- A.3 Subtitle acquisition, provenance, and recovery
- A.4 Complete-coverage optimization over languages and countries
- A.5 Multi-layer subtitle verification and manual audit
- A.6 Subtitle-grounded language-safety gold construction and audit
- A.7 Task-oriented label curation and normalization
- A.8 Final film-entity structure and field provenance
- A.9 Dataset statistics and coverage audits

B Evaluation Design and Reproducibility
- B.1 Model coverage and generation settings
- B.2 Metric definitions
- B.3 Task and evaluator instructions
- B.4 Independent narrative-judge agreement
- B.5 Response-format reliability
- B.6 Title/year recall and character-name counterfactuals

C Additional Evidence for RQ 1: Model Capability Profiles
- C.1 English and cross-lingual headline results
- C.2 Paired comparisons of leading composite scores
- C.3 Narrative failure patterns

D Additional Evidence for RQ 2: Cross-Lingual Behavior
- D.1 English-to-cross-lingual overview
- D.2 Narrative scores by subtitle language
- D.3 Genre stability and label-specific failures
- D.4 Paired cross-lingual uncertainty

E Additional Evidence for RQ 3: Age and National-Rating Calibration
- E.1 Age error magnitude and direction
- E.2 Country-level calibration and language sensitivity

F Additional Evidence for RQ 4: Auditable Language-Safety Assessment
- F.1 Safety-evidence errors by category and model

G Additional Analysis: Within-Family Scaling

H Qualitative Case Studies

I Limitations, Intended Use, and Data Statement

A Dataset Construction and Quality Assurance

A.1 Source Roles and Construction Overview

Construction sequence. Construction proceeds through six stages: source-inventory acquisition, cross-source entity resolution, task-metadata filtering, complete subtitle-language coverage selection, complete national-rating coverage selection, and subtitle verification before final record assembly. Tables 2 and 3 identify the source and retention rule at each stage. Automated subtitle flags initiate manual inspection, and unresolved structural or decoding failures prevent ingestion.

Source roles and provenance. Kids-in-Mind provides the starting inventory, Common Sense Media supplies age-suitability judgments, IMDb provides film metadata and the shared title identifier, and subtitle repositories provide the model input. Review metadata is retained for provenance and audit, while models receive subtitle text as their only film evidence.

Table 2: Source roles in CineSubBench construction. Each source contributes a distinct component of the benchmark, while the final record preserves provenance links and exposes only subtitle text to models during evaluation.

Source	Primary role	Quality and provenance treatment
Kids-in-Mind	Initial 6,186-film inventory; review-derived language metadata.	The native identifier supports cross-source traceability; records without a usable IMDb link for subtitle retrieval are excluded.
Common Sense Media	General age-suitability label, storyline, and review page.	Cross-linked through IMDb ID; anomalous embedded IMDb links receive case-by-case manual adjudication before retention.
IMDb	Stable title identifier; plot synopsis, runtime, film metadata, and regional motion-picture ratings.	Provides the entity key used to link sources and subtitle assets; populated plot-synopsis and rating fields are mandatory for the metadata-complete cohort.
Subtitle providers	Timestamped SRT files in six languages, with provider metadata.	OpenSubtitles is the primary source and SubDL supports recovery. The highest-download candidate is selected per film and language when alternatives exist; its count is retained in the JSON. Structural and temporal verification remains independent of this popularity proxy.

A.2 Cross-Source Entity Resolution and Progressive Filtering

1. **Begin with Kids-in-Mind.** The initial scrape produced 6,186 films. We use the IMDb title identifier to query subtitles and join the other sources. For example, `tt0465602` identifies an IMDb title independently of how its title or year is written on a review page.
2. **Validate identifiers.** Among 4,307 films with IMDb links, 33 had malformed links or lacked usable subtitle-search parameters. The remaining 4,274 formed the subtitle-acquisition baseline. A shared title and release year alone did not establish a match, because remakes and namesakes can collide.
3. **Link Common Sense Media.** We queried it for films in that baseline and matched through IMDb IDs. An anomalous page was reviewed manually: if its embedded IMDb link pointed to a different film, the page was not accepted by URL presence alone. The resolved intersection contained 2,831 films.

4. **Require IMDb task metadata.** We extracted metadata for those linked films and retained records with populated plot-synopsis and motion-picture-rating fields. This left 2,322 candidates for coverage selection.

Table 3: Progressive construction gates in CineSubBench. Counts report the number of films retained after each filtering or coverage decision. The detailed subtitle-verification audit is applied to the resulting 1,231-film multilingual cohort.

Step	Decision	Retention rule	Films
1	Source crawl	Kids-in-Mind films collected	6,186
2	IMDb link screen	4,307 films with links; exclude 33 invalid/search-unusable entries	4,274
3	Entity linking	IMDb-ID agreement across Kids-in-Mind and Common Sense Media; anomalous links manually reviewed	2,831
4	Metadata completeness	Populated IMDb plot synopsis and motion-picture-rating fields	2,322
5	Language coverage	Subtitles in all six selected languages	1,231
6	Country coverage	Ratings from all ten selected countries in the final benchmark	1,012

A.3 Subtitle Acquisition, Provenance, and Recovery

Subtitle collection begins from the 4,274-film identifier-validated baseline, while the full verification protocol in Table 5 is applied to the selected 1,231-film multilingual cohort. The acquisition and recovery procedure is:

1. **Retrieve and organize.** We retrieved SRT files through the OpenSubtitles Python interface using IMDb identifiers and organized them by film identifier and subtitle language. This identifier-based structure preserved the cross-source entity links established in the preceding stage.
2. **Choose among candidate files.** For each film and subtitle language with multiple available SRT files, we chose the candidate with the highest download count reported by the source website. Download count serves only as the initial selection heuristic. The selected file's `download_count` is retained in the JSON, and every chosen track remains subject to structural, language-consistency, and temporal verification that can trigger manual review or replacement.
3. **Recover suspect tracks.** Structural, encoding, language-consistency, or timing concerns led to alternative downloads through SubDL and manual checks of SubDL and OpenSubtitles web pages. A curator inspected every track flagged by the automated verification thresholds before deciding whether the anomaly reflected a genuine asset problem.
4. **Correct the source asset when possible.** A repairable millisecond separator, such as `00:09:09:949`, could be changed to the SRT form `00:09:09,949`; dialogue content was never rewritten as part of this repair. When multiple providers showed the same upstream limitation, the best available native asset could be retained after curator review. Unresolved parsing or decoding failures prevented ingestion.

A.4 Complete-Coverage Optimization over Languages and Countries

Complete coverage is the mechanism that makes the MultiX comparisons matched: every retained film is evaluated over the same six subtitle languages and, after country selection, the same ten national rating systems.

Choosing subtitle languages. We first apply Algorithm 1 to the 2,322-film metadata-complete cohort. The availability histogram in Figure 6a defines a pool of the 29 most frequent subtitle languages; the 29th, Czech, appears for 1,205 films. For a candidate set S, let $M(S)$ be the films with an SRT file in *every* language in S. We score S by $|S|\,|M(S)|$, the number of subtitle files retained under common film coverage.

1. **Frequency-prefix baseline (Approach A).** For each set size K, we evaluate the top K languages in the frequency ranking. Its best result has five languages and 1,458 films, giving $1{,}458 \times 5 = 7{,}290$ files. The top-six prefix has 1,157 films and 6,942 files.
2. **Combination search (Approach B).** For each K, compare candidate K-language subsets from the same pool using complete film coverage. Its best five-language set equals the prefix result. At six languages, a different set retains 1,231 films and $1{,}231 \times 6 = 7{,}386$ files. This is 444 more files than the six-language prefix and 96 more than the best prefix configuration.
3. **Selected set.** Arabic (ar), English (en), Indonesian (id), Persian (fa), Romanian (ro), and Vietnamese (vi) form the six-language result. The selected set includes a language ranked eighth in marginal frequency, illustrating why selecting languages one at a time by popularity can miss the best shared-coverage set. At $K \geq 7$, complete coverage falls below 1,000 films.

Choosing rating countries. Starting with those 1,231 films, we first remove a film marked `Unrated` or `Not Rated` only if that is its sole motion-picture-rating entry. For example, a film with an unrated tag *and* a valid United Kingdom rating remains eligible. We then apply the same complete-coverage search in Algorithm 1, substituting national rating systems for subtitle languages. The candidate pool comprises the 21 most frequent country systems shown in Figure 6b, and a film belongs to $M(S)$ only when it has a valid rating from every country in S. A ten-country set retains approximately one thousand films while supporting a broad national comparison. The selected countries are Australia, Brazil, France, Germany, the Netherlands, Singapore, South Korea, Sweden, the United Kingdom, and the United States. The final benchmark contains 1,012 films with complete ratings in these ten systems.

Table 4: Country-specific ordered motion-picture rating label spaces used in CineSubBench. Labels are normalized within each national classification system and retained as distinct ordinal scales rather than treated as interchangeable across countries.

Country	Ordered rating labels
Australia	G, PG, M, MA15+, R18+
Brazil	Livre, 10, 12, 14, 16, 18
France	Tous publics, 12, 16, 18
Germany	0, 6, 12, 16, 18
Netherlands	AL, 6, 9, 12, 14, 16, 18
Singapore	G, PG, PG13, NC16, M18, R21
South Korea	All, 12, 15, 19
Sweden	Btl, 7, 11, 15
United Kingdom	U, PG, 12, 15, 18
United States	G, PG, PG-13, R, NC-17

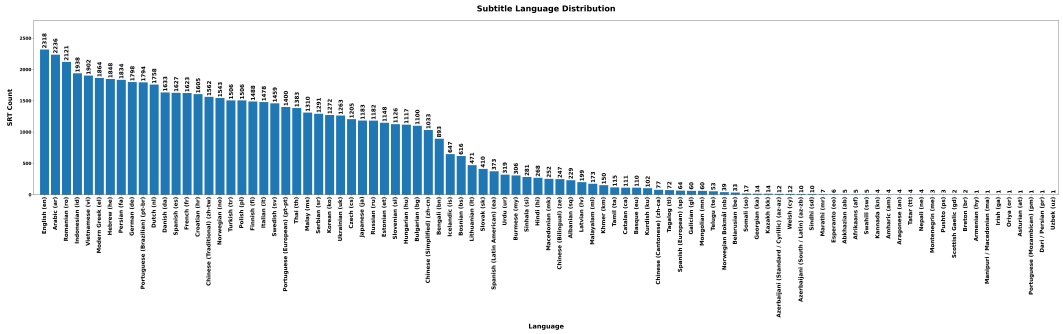
Algorithm 1: Complete-coverage set selection

Input: film-by-attribute availability map A, candidate pool $\mathcal{X}$, maximum cardinality $K_{\max}$
Output: selected attribute set $S^\star$ and complete-coverage film set $M(S^\star)$

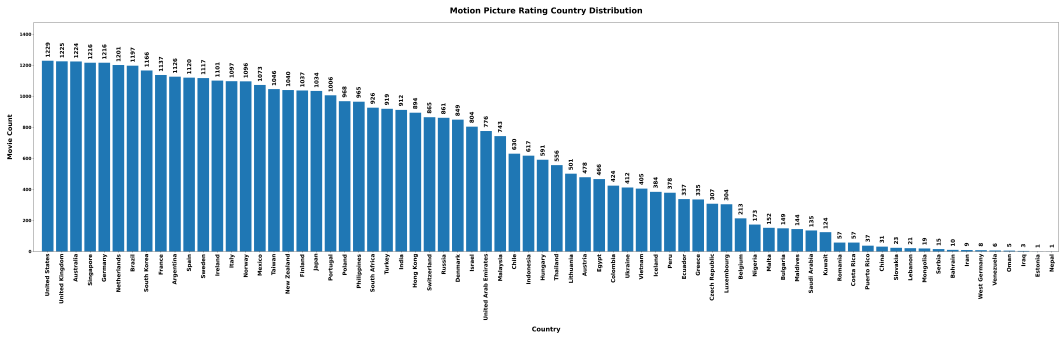
1. Set $q^\star \leftarrow -\infty$.
2. For each $K \in \{1, \ldots, K_{\max}\}$, enumerate every $S \subseteq \mathcal{X}$ with $|S| = K$.
3. For each S, compute $M(S) \leftarrow \{i : A_{i,x} = 1 \; \forall \, x \in S\}$ and $q \leftarrow |S| \times |M(S)|$.
4. If $q > q^\star$, set $(S^\star, M(S^\star), q^\star) \leftarrow (S, M(S), q)$.
5. Return $S^\star$ and $M(S^\star)$.

A.5 Multi-Layer Subtitle Verification and Manual Audit

The verification layers in Table 5 target distinct failure modes in the selected 1,231-film multilingual cohort: a provider can return an error page instead of dialogue, a syntactically valid track can cover only part of a film, and an otherwise usable track can contain malformed timestamps. Threshold crossings trigger manual review, after which the curator determines whether the track should be retained, repaired, or replaced.



(a) Subtitle-language availability



(b) National rating-system availability

Figure 6: Availability before complete-coverage optimization. Algorithm 1 is applied to (a) the top 29 subtitle languages over the 2,322 metadata-complete films and (b) the top 21 national rating systems over the 1,231-film multilingual cohort.

Table 5: Subtitle-level verification layers applied to the optimized six-language cohort. Each automated trigger initiates targeted manual inspection, source recovery, or an ingestion stop, rather than silently removing the affected subtitle track.

Verification layer	Automated rule	Audit purpose and response
Provider download count	Among alternative SRT files for the same film and language, select the file with the highest source-reported count; retain the selected count in the JSON.	Use popularity only as an initial quality proxy. Subject the selected file to the remaining checks and recover it if verification finds a genuine problem.
Provider-body integrity	Scan subtitle bodies for server-error signatures, including Cloudflare and Guru Meditation messages.	Detect error pages returned in place of dialogue; retrieve or inspect a replacement asset.
Timestamp syntax	Require valid SRT time ranges with hours, minutes, seconds, millisecond separator, and arrow structure (e.g., `00:09:09,949 --> 00:09:12,179`).	Catch malformed headers before JSON creation; repair or re-download the source asset.
Chronological consistency	Compare the start and end time of every subtitle block.	Flag negative-duration temporal inversions.
Single-block lifespan	Flag any subtitle block lasting 15 minutes or more.	Detect frozen or malformed subtitle segments.
Delayed onset	Flag a first operational subtitle beginning after five minutes.	Detect missing opening dialogue or late-start tracks.
Runtime coverage	Flag a final subtitle end time divided by IMDb movie runtime below two-thirds.	Detect tracks that appear to stop before later acts of the film.
Macro-intervals	Flag a 15-minute or larger gap between adjacent subtitle blocks.	Detect missing internal narrative segments or dropped blocks.

Examples of audit decisions.

- **False subtitle download.** If the file body contains a Cloudflare or "Guru Meditation" error message, the payload is a provider response rather than film dialogue; the curator seeks a replacement.
- **Short temporal coverage.** For an illustrative 120-minute film, a track whose last subtitle ends at 70 minutes has endpoint coverage $70/120 = 0.58$, below the $2/3$ trigger. Its source and timeline are inspected before it is accepted or replaced. A silent ending alone is not assumed to be an error.
- **Potential missing span.** A first subtitle after five minutes or an internal gap of at least 15 minutes raises a flag. A wordless opening or long action sequence may explain the gap, so a curator checks the track before deciding.
- **Malformed block.** A subtitle block whose end precedes its start, or one that stays active for at least 15 minutes, is examined against the source file. Correctable time formatting is repaired there; an unresolved block prevents structured ingestion.

A.5.1 Structural and Encoding Safeguards

Strict structural ingestion. Every expected SRT block has three components: a sequential index, a timestamp header, and a text payload. A combined splitter first separates candidate headers and bodies; a header parser extracts the index and time range only when the header satisfies the SRT structure. Thus a file containing plausible dialogue but a malformed header is not partially ingested. A non-matching header is recorded as a structural failure and stops ingestion until the source asset is manually corrected or cleanly re-downloaded. This prevents malformed blocks from being silently omitted.

Deterministic decoding. Raw subtitle files do not share a universal character encoding, especially across Arabic, Persian, and European-language sources. We therefore attempt the ordered sequence in Table 6, without suppressing undecodable bytes. If every tier fails, the asset is rejected until manual normalization or a clean replacement is available. A successful byte-level decode does not itself prove that the text is in the expected language; suspicious language or character content is checked during manual review.

Table 6: Deterministic subtitle-decoding cascade. Decoding proceeds to the next encoding tier only when the preceding one fails to read the source asset; lossy error suppression is never used.

Tier	Encoding	Purpose and representative case
1	`utf-8-sig`	Default modern-web decoding while removing a byte-order mark when present.
2	`utf-16`	Handles UTF-16 subtitle assets, including files exported by automated transcription or speech-to-text systems.
3	`cp1256`	Handles legacy Windows Arabic and Persian encodings, preserving right-to-left dialogue instead of dropping unsupported characters.
4	`cp1252`	Handles legacy Western-European encodings and extended diacritics, such as é, ü, and ñ.
Failure	none succeeds	Reject the asset and require manual normalization or a clean replacement before ingestion resumes.

A.5.2 Subtitle Sanitization and Artifact Removal

As part of data preparation, we apply an automated, auditable sanitization procedure to all six subtitle-language tracks. Across 8,133,088 subtitle entries from 1,012 films, the procedure removes 6,594 entries containing non-narrative distribution artifacts, including translator or subtitle-team credits, personal contact details, social-media handles, subtitle-provider advertisements, download promotions, and external promotional links. These elements originate from subtitle-distribution files rather than the films' narrative content.

To preserve meaningful content, entries containing contact information are not removed indiscriminately. In 65 cases where an email address occurs within an otherwise narrative-relevant subtitle entry, such as an on-screen message, the surrounding text is retained and only the address is replaced with `[email redacted]`. Sanitization preserves film identifiers, timestamps, language-track metadata, and original subtitle indices without renumbering, maintaining temporal alignment and links between language-content annotations and their supporting subtitle evidence.

A.5.3 Final Dataset Consistency Checks

After source recovery and JSON creation, the integrated records were checked for:

- unique, stable film identifiers and source links that point to the corresponding film;
- one named subtitle track for each of the six languages, with indexed entries and start/end timecodes;
- ten country-rating entries per film in the defined country order, with one normalized movie-rating label per country;
- populated task fields, consistent field types, and conformance to the dataset schema; and
- subtitle-grounded language evidence whose indices resolve to actual English subtitle entries.

These checks are paired with manual review of narrative fields, rating normalization, and discrepancies between subtitle-derived language counts and Kids-in-Mind summaries; Appendix A.6 details the latter audit.

A.6 Subtitle-Grounded Language-Safety Gold Construction and Audit

The annotation workflow connects the review-level counts in Kids-in-Mind to auditable, line-level evidence in the English subtitle track:

1. **Collect source counts and examples.** For each film, we retained the prose language assessment in `languageContentReference`, including its reported category counts and parenthetical examples. These review-level counts provide targets for comparison but no subtitle locations; the nested `source` field preserves Kids-in-Mind provenance.
2. **Operationalize the glossary.** We followed the definitions and examples in the Kids-in-Mind glossary, together with recurring examples in the film summaries, to construct category-specific wordlists and regular expressions. The benchmark schema keeps these external counts in `languageContentReference` and the audited subtitle-grounded labels in `languageContentAssessment`.
3. **Normalize and scan the subtitles.** Before matching, we remove HTML-style markup and bracketed or all-capital stage directions. We then scan every English subtitle entry, counting all lexical occurrences and recording the index of each line containing at least one match. Thus occurrence count and evidence-line count remain distinct.
4. **Audit discrepancies.** We compare extracted counts with the parsed Kids-in-Mind counts, screen the resulting discrepancy reports, and inspect flagged and large disagreements in subtitle context. We revise a rule only for a broad, defensible lexical gap or false-positive pattern; every adjustment is manually verified against retrieved lines.
5. **Preserve evidence-based disagreement.** We do not force the two sources to agree. A difference may reflect a different subtitle version, a censored bleep, a gesture, or reviewed content that is not lexicalized in the evaluated subtitle track. In such cases, the subtitle-derived result is retained.
6. **Finalize auditable gold.** The audited subtitle occurrences and evidence indices form the benchmark gold labels. The original Kids-in-Mind summary remains in the record as provenance and an independent audit reference.

Operational categories. The policy covers **strong profanity** (F-word forms and close derivatives), **crude bodily language** (scatological and crude anatomical language), **mild obscenity** (lower-intensity obscene or impolite expressions), and **religious profanity/exclamation** (religious profanity and exclamatory uses, excluding ordinary religious dialogue). Items not recoverable from subtitle

text, such as obscene hand gestures, are excluded. Broad derogatory or name-calling categories are likewise excluded because the present rules do not support sufficiently precise and reproducible boundaries.

Stored evidence and auditability. For each category, the dataset stores the `occurrenceCount`, `evidenceLineCount`, and `evidenceSubtitleIndices`. The indexed English subtitle entries retain the text and start/end timecodes, so every gold decision can be recovered and inspected without duplicating offensive text in the label field. A separate diagnostic file retains matched spans and category information for reproducibility and future rule refinement. Three malformed source rating values caused by title-number leakage were reconstructed from their structured full-rating fields before that metadata was used for audit.

Examples of audit-informed decisions. The mismatch audit identified apostrophe variants such as shortened F-word forms, which were added because they are explicit subtitle evidence. It also identified ordinary uses that must not be counted: a capitalized personal name, food-related uses of "breast," spatial uses such as "bottom of," and ordinary religious dialogue are filtered by contextual rules. Conversely, broad wildcard patterns for partially censored words were rejected after producing unrelated matches, and obscene gestures remain excluded because they cannot be recovered from subtitle text. These decisions make the annotation reproducible without mechanically reproducing a review site's count.

A.7 Task-Oriented Label Curation and Normalization

Task selection and complete coverage. We retained core tasks only when their reference labels could be provided uniformly across the final benchmark and could be meaningfully evaluated from subtitle evidence. This criterion excluded otherwise plausible fields with incomplete coverage. For example, tagline generation was removed because taglines were available for only 946 of the final 1,012 films. The key-message task was retained because its references were made complete across the final cohort and because it captures thematic information that can be inferred from dialogue and narrative progression.

Narrative-reference cleaning. Plot, synopsis, and storyline fields were audited for material external to the film narrative, including production history, remake notes, franchise context, and other database-specific editorial text. When such material appeared, the reference was cleaned to retain film-internal narrative content. Raw source text was retained in internal curation records for traceability. A small number of missing key-message entries were completed through curator review so that the final task has full reference coverage.

Age suitability and national-rating normalization. General age suitability and formal motion-picture classification are represented as separate targets. Common Sense Media ratings are retained as family-oriented age-suitability labels. National motion-picture ratings are normalized independently within each country's classification system using movie-rating labels rather than television, app-store, or platform-specific categories. Missing, obsolete, or multiple candidate values are resolved to one contemporary ordered label for the corresponding country. For example, South Korean adult ratings are normalized to the current 19+ category, while the United Kingdom's theatrical 12A and video 12 categories are represented by the benchmark label `12`. Table 4 lists the resulting ordered spaces.

The benchmark keeps these country-specific label spaces as separate ordinal systems. Continuous pseudo-normalized scores are excluded from the benchmark schema, avoiding artificial numerical equivalence between national classifications with different institutional meanings.

Task-level validation and manual audit. Curator review supplements the automated dataset checks in Appendix A.5. Narrative references are checked against their source text after cleaning; ambiguous or historical country-rating values are resolved within their national systems; and disagreements between external language-content summaries and subtitle-derived evidence are inspected in the corresponding subtitle context. Final schema validation confirms complete task fields and consistent field types for every retained film.

A.8 Final Film-Entity Structure and Field Provenance

After construction, verification, normalization, and annotation, CineSubBench is represented as a JSON array in which each object corresponds to one film. The schema keeps film metadata, task references, provenance, and six timestamped subtitle tracks in a consistent record. The types below describe the benchmark schema.

CineSubBench Film Entity (JSON Object)

Identity and film metadata

`id` *(Integer)* — Unique sequential identifier for the film.
`title` *(String)* — Canonical film title.
`releaseYear` *(String)* — Year in which the film was released.
`duration` *(String)* — Human-readable film runtime.
`imdbRating` *(String)* — IMDb user-rating value recorded during collection.
`countriesOfOrigin` *(String)* — Country or countries associated with the production.
`originalLanguages` *(String)* — Film's original spoken language or languages.
`directors` *(Array of Strings)* — Credited director names.
`writers` *(Array of Strings)* — Credited writer names.
`topCast` *(Array of Objects)* — Principal cast entries, each containing an `actor` and primary `character`.
`fullCast` *(Array of Objects)* — Expanded cast entries with `actor`, `characters`, and any `additionalTexts`.

Narrative and classification references

`plot` *(String)* — Concise premise-level narrative reference.
`synopsis` *(String)* — Longer, spoiler-aware account of the narrative arc.
`storyline` *(String)* — Short source-provided storyline description.
`keyMessage` *(String)* — Reference statement of the film's principal theme or lesson.
`genres` *(Array of Strings)* — Gold multi-label genre set.
`topics` *(String)* — Source-provided topical descriptors.
`ageSuitabilityRating` *(String)* — General age-suitability label, expressed as an age threshold such as `13+`.
`motionPictureRatings` *(Array of Objects)* — Ratings for the ten selected national systems. Each entry stores `country` (String), `label` (String), `minimumAge` (Integer), and `ordinal` (Integer).

Language-content assessment

`languageContentReference` *(Object)* — External review information retained for provenance and audit: `rating` (Integer), `summary` (String), and `source` (String).
`languageContentAssessment` *(Object)* — Audited labels derived from the English subtitle track. It records `inputLanguage` (String) and a `categories` object containing `strongProfanity`, `crudeBodilyLanguage`, `mildObscenity`, and `religiousProfanityAndExclamation`.
`occurrenceCount` *(Integer)* — Number of matched expressions for one language-content category.
`evidenceLineCount` *(Integer)* — Number of distinct subtitle entries containing category evidence.
`evidenceSubtitleIndices` *(Array of Integers)* — Indices that resolve each category's evidence to the English subtitle entries.

Provenance and multilingual subtitle input

`sources` *(Object)* — Source URLs stored as `imdbLink`, `commonSenseLink`, and `kidsInMindLink`.
`subtitles` *(Array of Objects)* — Six subtitle tracks, one per benchmark language. Each track contains `language` (String), `downloadCount` (Integer), and `entries` (Array of Objects).
`entries` *(Array of Objects)* — Ordered subtitle units containing `index` (Integer), `timeframe` (String), and `content` (String).

Table 7 makes the field-level provenance explicit. The mapping names the originating website rather than the later cleaning or normalization stage applied during benchmark construction.

Table 7: Mapping from public CineSubBench fields to their source or construction process. Website names link to the corresponding source. The subtitle-derived `languageContentAssessment` follows Kids-in-Mind definitions and uses its reference counts for auditing, while the final labels and evidence are constructed directly from the English subtitle track.

Source website	Public dataset fields
Common Sense Media	`ageSuitabilityRating`; `storyline`.
Kids-in-Mind	`keyMessage`; `languageContentReference`.
IMDb	`title`; `releaseYear`; `duration`; `imdbRating`; `countriesOfOrigin`; `originalLanguages`; `genres`; `topics`; `plot`; `synopsis`; `motionPictureRatings`; `directors`; `writers`; `topCast`; `fullCast`.
OpenSubtitles and SubDL	`subtitles`, including each track's `language`, `downloadCount`, and timestamped `entries` (`index`, `timeframe`, and `content`).
CineSubBench annotation pipeline	`languageContentAssessment`, containing manually audited, subtitle-localized category counts and evidence indices derived from the English subtitle track.

Schema note. `id` is assigned by CineSubBench, while `sources` stores the IMDb, Common Sense Media, and Kids-in-Mind provenance URLs for each film.

A.9 Dataset Statistics and Coverage Audits

The tables and figures in this section describe all 1,012 films in the benchmark. Core counts are presented compactly, while task distributions and temporal-coverage diagnostics receive full-width treatment.

Table 10 reports reference-output lengths for narrative tasks; Tables 13, 14, and 11 report the supporting label distributions.

Table 8: Core CineSubBench dataset statistics.

Statistic	Value
Films	1,012
Subtitle tracks	6,072
Subtitle entries	8.13M
Release years	1977–2026
Runtime, mean / median	114.6 / 113.0 min
Runtime, 95th pct. / max	149.4 / 242.0 min
IMDb rating, mean / median	6.80 / 6.80
IMDb rating, min / max	3.5 / 8.8
Raw genre labels	149
Core genre labels	22
Raw genres per film, mean / median	4.08 / 4.0

A.9.1 Subtitle-Duration Coverage Audit

We audit subtitle duration and density relative to the corresponding film runtime and English track. The audit measures temporal span/runtime, subtitle-interval density/runtime, token ratio to English, and entry ratio to English. Tracks are flagged when span/runtime is below 0.65, token ratio is below 0.60, entry ratio is below 0.70, or temporal span differs unusually from English. Because subtitle activity need not span every visual moment, these thresholds initiate inspection rather than automatic deletion.

Arabic duration and density audit. Table 12 reports a mean Arabic temporal span/runtime of 0.931 and a fifth percentile of 0.868, with no track below the short-runtime threshold. Arabic nevertheless has lower subtitle density than English: 27 tracks fall below the token-ratio threshold and 120 below the entry-ratio threshold. A supplementary correlation audit finds little association between these density indicators and Arabic narrative quality, indicating that duration-level truncation alone does not account for the observed model gap.

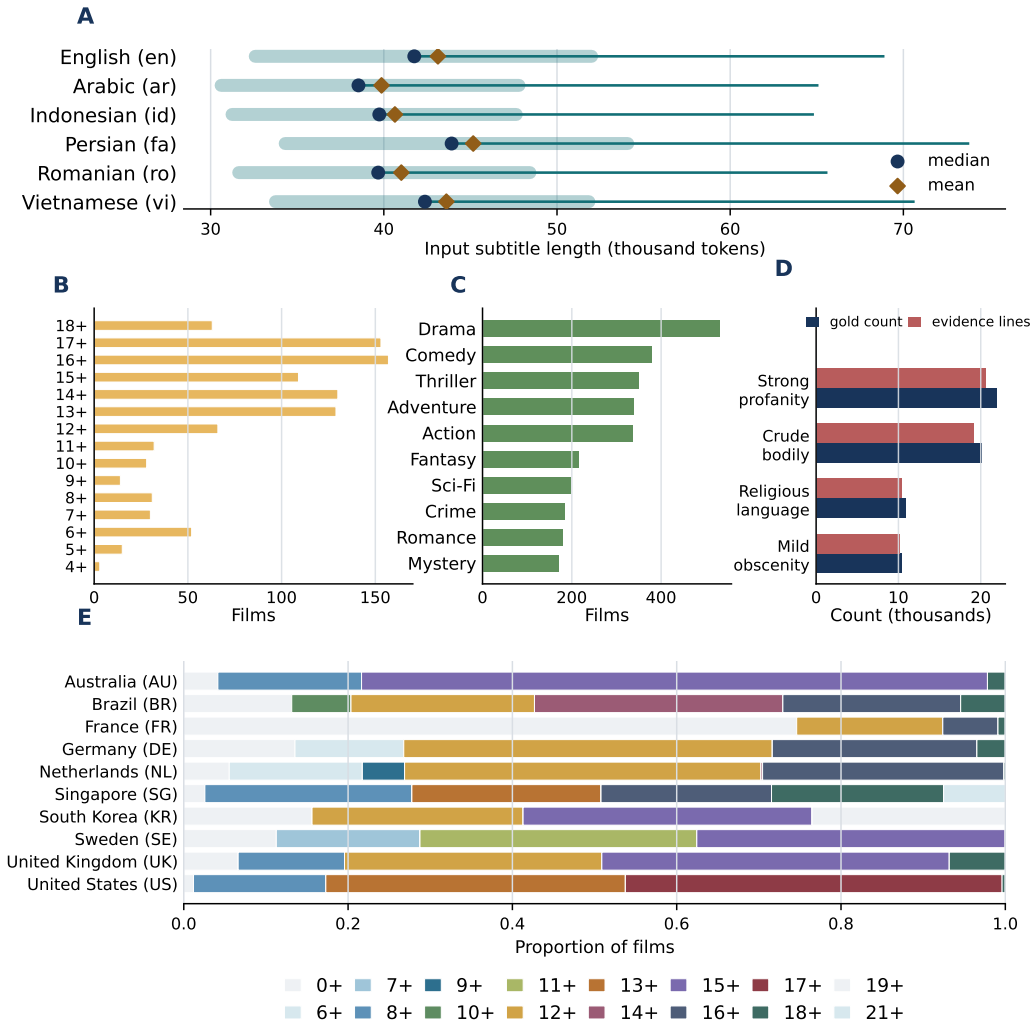


Figure 7: Task-relevant statistics for all 1,012 CineSubBench films. Panels show (A) subtitle input length by language, (B) age-suitability label distributions, (C) normalized genre frequencies, (D) subtitle-grounded language-safety counts and corresponding evidence-line counts, and (E) national motion-picture rating distributions represented by minimum-age values. Token counts are computed from prompt-formatted subtitle inputs using the `o200k_base` tokenizer.

Table 9: Subtitle input size by language. Token counts are computed from prompt-formatted subtitle inputs using the tokenizer adopted for prompt-length accounting; they reflect tokenization-dependent input length rather than language-neutral semantic length.

Language	Mean tok.	Median tok.	95th pct.	Max tok.	Median entries	Median ratio
English (en)	43,123	41,759	68,833	144,268	1,418	1.00
Arabic (ar)	39,867	38,536	65,023	98,701	1,214	0.95
Indonesian (id)	40,641	39,739	64,760	120,034	1,308	0.97
Persian (fa)	45,162	43,916	73,731	112,920	1,325	1.08
Romanian (ro)	41,010	39,676	65,542	129,799	1,244	0.99
Vietnamese (vi)	43,613	42,374	70,575	108,761	1,322	1.05

Table 10: Reference output length statistics for narrative tasks.

Reference field	Mean	Median	P95	Max
Plot	33	33	50	66
Synopsis	1,448	1,262	2,802	7,151
Storyline	156	154	242	437
Key message	14	12	26	55

Table 11: Subtitle-grounded language-content statistics. Evidence is defined over the English subtitle track.

Category	Gold count	Evidence lines
Strong profanity	21,892	20,602
Crude bodily language	20,131	19,149
Religious profanity/exclamation	10,861	10,447
Mild obscenity	10,360	10,088

Table 12: Subtitle-duration coverage audit by language. Span/runtime measures the fraction of the film timeline covered by each subtitle track, while token and entry ratios compare each track with the English subtitle track for the same film. Low-token and low-entry counts use thresholds of < 0.60 and < 0.70 relative to English, respectively.

Language	Flagged	Span/run.	Span p05	Tok. ratio	Entry ratio	Low tok.	Low entry
English (en)	9	0.95	0.89	1.00	1.00	0	0
Arabic (ar)	142	0.93	0.87	0.95	0.88	27	120
Indonesian (id)	112	0.95	0.89	0.97	0.93	28	84
Persian (fa)	123	0.96	0.88	1.08	0.95	23	77
Romanian (ro)	155	0.95	0.89	0.98	0.89	27	125
Vietnamese (vi)	93	0.96	0.90	1.04	0.95	20	57

Table 13: Most frequent normalized core genres.

Genre	Films
Drama	531
Comedy	380
Thriller	350
Adventure	340
Action	336
Fantasy	215
Sci-Fi	199
Crime	185
Romance	181
Mystery	171
Family	158
Animation	135
Horror	114
Biography	89
Music	78

Table 14: Common Sense age-suitability labels.

Age label	Films
16+	157
17+	153
14+	130
13+	129
15+	109
12+	66
18+	63
6+	52
11+	32
8+	31
7+	30
10+	28
5+	15
9+	14
4+	3

B Evaluation Design and Reproducibility

B.1 Model Coverage and Generation Settings

Table 15 reports model access, subtitle languages, and task coverage for every comparison. The main six-language evaluation contains nine models spanning closed frontier, closed baseline, and open-weight models. A separate Ministral 3B, 8B, and 14B experiment provides a controlled English-only within-family scaling comparison. GPT-5.6 Luna serves as the common reference-based evaluator for narrative generation. Narrative, genre, age-suitability, and country-rating analyses use the matched six-language film set; subtitle-grounded language-safety assessment remains English-only because its gold evidence is indexed to the English track.

All API models use deterministic decoding with temperature 0, while local models use greedy decoding. We apply the same task-specific output ceilings across providers and validate responses against the corresponding task schema. Table 16 gives the complete settings and provider parameter mappings.

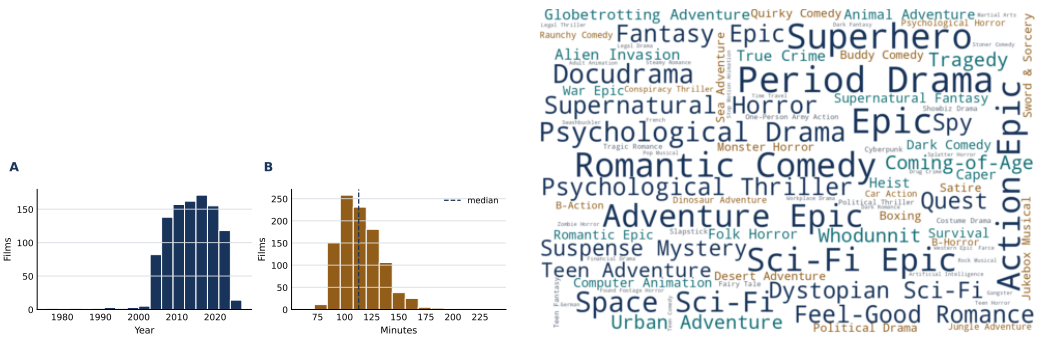
Table 15: Models and task coverage in CineSubBench. All models receive subtitle text as their only film evidence and generate English outputs for narrative generation tasks. Six-language comparisons use the same matched film set, while evidence-grounded language-safety assessment is evaluated only on the English subtitle track.

Evaluation role	Models	Model access	Subtitle inputs	Tasks
Closed frontier	GPT-5.6 Sol; Gemini 3.8 Flash; Claude Haiku 4.5	OpenAI, Gemini, and Anthropic APIs	English + Arabic, Indonesian, Persian, Romanian, Vietnamese	Narrative and cultural; English language safety where available
Closed baseline	GPT-5 Nano; Gemini 3.5 Flash Lite	OpenAI and Gemini APIs	English + Arabic, Indonesian, Persian, Romanian, Vietnamese	Narrative and cultural; English language safety where available
Open-weight	DeepSeek V4 Pro 1.6T; Qwen3 235B; Mistral 4 119B; Llama 4 Scout 17B	OpenRouter API	English + Arabic, Indonesian, Persian, Romanian, Vietnamese	Narrative and cultural; language safety for completed English lanes
Local scaling	Ministral 3B, 8B, and 14B	Local GPU inference	English	Controlled English scaling analysis
Narrative evaluator	GPT-5.6 Luna	OpenAI API	Narrative outputs from all compared models	Reference-based narrative scoring and structured error annotations

B.2 Metric Definitions

The main paper reports one headline measure for each task; this section provides the complete definitions and secondary diagnostics. Within each comparison, all metrics are computed over the same evaluated films.

Narrative generation. Plot, synopsis, and key-message generations admit substantial valid paraphrase, making surface-overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) poorly suited to capturing narrative quality; prior work likewise finds that automatic metrics can correlate weakly with human judgments for open-ended generation (Fabbri et al., 2021; Zhang & Bansal, 2021; Liu et al., 2023). Human evaluation provides richer semantic judgments but is costly, time-consuming, and difficult to scale across the benchmark's models, tasks, languages, and long-form inputs (Zhang & Bansal, 2021). We therefore use GPT-5.6 Luna as a reference-based



(a) Release year and runtime **(b) Topical coverage**

Figure 8: Additional corpus context. (a) Release-year and runtime coverage. (b) Topic frequency after splitting comma-separated tags and removing generic core-genre labels; font size indicates frequency.

Table 16: Standardized decoding settings across model families. API models use deterministic decoding with temperature 0, while local models use greedy decoding. Task-specific output ceilings are shared across providers: 1,200 tokens for narrative generation, 1,400 for cultural prediction, 2,500 for LC Assessment, and 4,096 for narrative judging. Provider adapters map these settings to the corresponding API parameters (e.g., `max_completion_tokens`, `max_output_tokens`, `max_tokens`, or `max_new_tokens`) without changing the numerical limits. All outputs are validated against the corresponding task schema.

Model access	Evaluated models	Decoding	Task-specific output ceilings
OpenAI API	GPT-5.6 Sol, GPT-5 Nano; GPT-5.6 Luna as narrative evaluator	Temperature 0	1,200 / 1,400 / 2,500 / 4,096
Gemini API	Gemini 3.8 Flash, Gemini 3.5 Flash Lite	Temperature 0	1,200 / 1,400 / 2,500 / 4,096
Anthropic API	Claude Haiku 4.5	Temperature 0	1,200 / 1,400 / 2,500 / 4,096
OpenRouter API	DeepSeek V4 Pro 1.6T, Qwen3 235B, Mistral 4 119B, Llama 4 Scout 17B	Temperature 0	1,200 / 1,400 / 2,500 / 4,096
Local inference	Ministral 3B, 8B, and 14B	Greedy	1,200 / 1,400 / 2,500 / 4,096

LLM judge with an explicit task-specific rubric, following recent work on structured model-based evaluation (Liu et al., 2023; Li et al., 2025; Gu et al., 2026). The judge first identifies concrete narrative errors and then assigns a score. To ensure comparability, every model is evaluated with the same judge, references, rubric, input ordering, and structured response format. Complete judge instructions appear in Appendix B.3.

For film i and narrative task $t \in \{\text{plot}, \text{synopsis}, \text{key}\}$, let $s_{i,t} \in \{1, \ldots, 5\}$ be the evaluator's score. The aggregate narrative score is

$$S_{\text{nar}} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \frac{s_{i,\text{plot}} + s_{i,\text{synopsis}} + s_{i,\text{key}}}{3}. \tag{2}$$

We additionally report plot, synopsis, and key-message scores separately and analyze the evaluator's task-specific error annotations.

Genre prediction. Genre prediction is multi-label. For true label set G_i and predicted set $\hat{G}_i$, the headline metric is micro-F1:

$$\text{F1}_{\text{micro}} = \frac{2 \sum_\ell \text{TP}_\ell}{2 \sum_\ell \text{TP}_\ell + \sum_\ell \text{FP}_\ell + \sum_\ell \text{FN}_\ell}. \tag{3}$$

We additionally report macro-F1, exact set match, and mean Jaccard similarity. Exact set match requires the complete predicted genre set to equal the gold set, while Jaccard similarity measures partial overlap.

Age suitability. Age labels are converted to their minimum ages in years. For gold age a_i and predicted age $\hat{a}_i$, mean absolute error is

$$\mathrm{MAE}_{\mathrm{age}} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} |\hat{a}_i - a_i|. \tag{4}$$

We also report exact match, accuracy within one year, and accuracy within two years. The error analysis further separates underprediction from overprediction.

Country-specific motion-picture ratings. Each country c retains its own ordered label space $\mathcal{R}_c$. For gold rating $r_{i,c}$ and predicted rating $\hat{r}_{i,c}$, country exact match is

$$\mathrm{EM}_{\mathrm{country}} = \frac{1}{|\mathcal{D}||\mathcal{C}|} \sum_{i \in \mathcal{D}} \sum_{c \in \mathcal{C}} \mathbb{1}\{\hat{r}_{i,c} = r_{i,c}\}. \tag{5}$$

We additionally report country-specific exact match, ordinal MAE, valid-prediction percentage, and performance relative to each country's majority-label baseline. Ordinal error is computed within each country's ordered label space rather than across a shared global scale.

Subtitle-grounded language safety. For film i and lexical category k, let $E_{i,k}$ and $\hat{E}_{i,k}$ denote the gold and predicted subtitle-index evidence sets. Strict-category evidence precision and recall are

$$P_{\mathrm{strict}} = \frac{\sum_{i,k} |\hat{E}_{i,k} \cap E_{i,k}|}{\sum_{i,k} |\hat{E}_{i,k}|}, \qquad R_{\mathrm{strict}} = \frac{\sum_{i,k} |\hat{E}_{i,k} \cap E_{i,k}|}{\sum_{i,k} |E_{i,k}|}, \tag{6}$$

with

$$F1_{\mathrm{strict}} = \frac{2 P_{\mathrm{strict}} R_{\mathrm{strict}}}{P_{\mathrm{strict}} + R_{\mathrm{strict}}}. \tag{7}$$

Strict matching requires both the correct subtitle index and the correct lexical category. Category-agnostic evidence metrics pool evidence indices across categories before matching, separating evidence localization from category assignment. We additionally report precision and recall separately, count MAE, per-category F1, gold-evidence miss rate, false-positive rate, and cross-category false-positive rate.

Cross-task composite. For models with complete comparable coverage, narrative, genre, age, and country metrics are first averaged equally over the six subtitle-input languages. Language-safety strict F1 remains English-only because the evidence annotations are defined over the English subtitle track. We normalize

$$N_{\mathrm{nar}} = 25(S_{\mathrm{nar}} - 1), \qquad N_{\mathrm{age}} = 100 \left(1 - \frac{\min(\mathrm{MAE}_{\mathrm{age}}, 16)}{16}\right), \tag{8}$$

where 16 is the span of the declared Common Sense age-label space from $2+$ to $18+$. The cultural component and overall score are

$$N_{\mathrm{cult}} = \frac{N_{\mathrm{age}} + \mathrm{EM}_{\mathrm{country}}}{2}, \qquad S_{\mathrm{overall}} = \frac{N_{\mathrm{nar}} + F1_{\mathrm{genre}} + N_{\mathrm{cult}} + F1_{\mathrm{LC}}}{4}. \tag{9}$$

The task-level tables retain every component in its original unit.

B.3 Task and Evaluator Instructions

The following boxes reproduce the model-facing task instructions and narrative-evaluation criteria used in the experiments. Line breaks separate input constraints, requested outputs, and scoring rules for readability without changing the instruction content.

Narrative Understanding and Generation Prompt

Identity. You are a professional film analyst, story editor, and metadata curator. You write concise, faithful film descriptions from subtitles only.

Task. Given subtitle entries for one film, produce all outputs for a narrative understanding and generation task group in one English JSON object: plot generation, synopsis generation, key-message generation, and genre prediction.

Input constraint. Use only the provided subtitles. The subtitles may be English or translated subtitles in another language, but every generated output must be written in English. Do not rely on outside knowledge, the film title, cast, franchise knowledge, reviews, or memorized plot information. If a detail is not inferable from the subtitles, omit it. The user message specifies the subtitle language as `inputLanguage`; treat that language label as authoritative.

Writing standards.

- **Plot (45–80 words):** a present-tense, third-person premise naming the central character or group, setup, main conflict, and driving situation.
- **Synopsis (180–300 words):** a present-tense, third-person, spoiler-aware account of the beginning, major developments, escalation, character goals or conflicts, and resolution when inferable. Connect events causally.
- **Key message (8–25 words):** one moral, social, emotional, or positive thematic takeaway, not another plot summary.
- **Genres (1–4 labels):** select exact strings from the allowed set using the narrative arc, characters, setting, tone, and dominant conflict; do not invent or substitute labels.

Allowed genre labels.

> Action · Adventure · Animation · Biography · Comedy · Crime · Documentary · Drama · Family · Fantasy · Film-Noir · History · Horror · Music · Musical · Mystery · Romance · Sci-Fi · Sport · Thriller · War · Western.

Output rules. Return only valid JSON that conforms to the provided schema. Do not include Markdown, commentary, confidence scores, explanations, wrapper keys, or snake-case aliases.

```
{
  "id": 0,
  "plot": "...",
  "synopsis": "...",
  "keyMessage": "...",
  "genres": ["Drama"]
}
```

Subtitle-Grounded Language-Safety Assessment Prompt

Identity. You are a subtitle evidence auditor for film language safety. You identify subtitle-grounded language-content categories using English subtitle lines.

Task framing. This is an exhaustive lexical extraction task, not a summary, representative-assessment, or severity-only task. Given English subtitle entries for one film or one non-overlapping subtitle chunk, scan all provided subtitle lines and return language-content counts and evidence indices in one compact JSON object. Long evidence lists are expected when many subtitle lines match.

Input constraint. Use only the provided English subtitles. Do not use the film title, cast, reviews, ratings databases, or outside knowledge. Do not output raw offensive wording. Evidence must be returned only as subtitle indices.

Category priority rules. Assign each matched expression to the most specific applicable category:

- `strongProfanity`: F-word forms and close derivatives.
- `crudeBodilyLanguage`: S-word forms, scatological terms, and crude anatomical language.
- `religiousProfanityAndExclamation`: religious profanity or exclamations.
- `mildObscenity`: lower-intensity obscene, impolite, softened, or euphemistic expressions; never a catch-all for stronger categories.

One subtitle line may appear under multiple categories if it contains multiple category types.

Counting and evidence rules. The category count is the number of matched expressions, not necessarily the number of evidence lines. Include every index containing a matching expression for that category. For example, two same-category expressions in one line count twice but yield one evidence index. If uncertain, prefer not to include an index.

Output rules. Return only valid JSON with flat schema keys. Do not include matched text, nested evidence objects, time intervals, Markdown, commentary, or wrapper keys.

Required flat JSON fields.

Count key	Evidence-index key
`strongProfanityCount`	`strongProfanityIndices`
`crudeBodilyLanguageCount`	`crudeBodilyLanguageIndices`
`mildObscenityCount`	`mildObscenityIndices`
`religiousProfanityAndExclamationCount`	`religiousProfanityAndExclamationIndices`

Reference-Based Narrative LLM-as-Judge Prompt Structure

Identity. The judge is instructed to act as a senior film critic, story analyst, and metadata quality auditor. The prompt uses a two-stage reference-based procedure to reduce impressionistic scoring.

Scope and inputs. For each sample, the judge receives the sample id, gold reference plot, gold reference synopsis, gold reference key message, and the model-generated plot, synopsis, and key message. The judge evaluates only the three narrative generation tasks. Genre prediction is excluded from the judge prompt and evaluated separately with automatic multi-label metrics.

Evaluation anchor. References are the evaluation anchor. The judge does not require exact wording, but penalizes missing central conflicts, missing resolutions, invented events, wrong relationships, weak causality, generic key messages, and outputs that are locally plausible but globally mis-centered. Each task is judged against its task-matched reference, while the other reference fields may be used only to disambiguate the film's central story or theme.
For plot scoring, the evaluator gives priority to the reference plot and uses the reference synopsis only to clarify factual context. It checks whether the response preserves the distinctive premise, central characters, setup, conflict, and dominant emphasis. A fluent but generic description does not receive a high score if it omits the film-specific trigger or shifts the story's center of gravity.

Stage 0: centrality audit. Before scoring, the judge internally identifies central characters or groups, relationships, motivations, conflicts, major events, turning points, end states, and dominant theme. This inventory is not output; it is used to avoid rewarding outputs that over-focus on secondary subplots or local subtitle-visible details.

Stage 1: standardized error identification. For each task, the judge first lists concrete errors with a type, severity (`minor`, `moderate`, or `major`), and short explanation. The type inventories are task-specific:

- **Plot:** `missingCentralConflict`, `missingProtagonistOrGroup`, `wrongSetup`, `wrongRelationship`, `inventedEvent`, `wrongEmphasis`, `tooVague`, `tooDetailed`, `styleOrLengthIssue`.
- **Synopsis:** `missingMajorEvent`, `missingResolution`, `wrongResolution`, `weakCausality`, `missingCharacterMotivation`, `wrongRelationship`, `inventedEvent`, `wrongEventOrder`, `tooVague`, `tooDetailed`, `styleOrLengthIssue`.
- **Key message:** `semanticMismatch`, `tooGeneric`, `tooPlotLike`, `missesPositiveMessage`, `contradictedByReference`, `unsupportedTheme`, `styleOrLengthIssue`.

Stage 2: score assignment. After identifying errors, the judge assigns dimension scores and an overall score on a 1–5 scale:

Score	Interpretation
5	Exceptionally faithful, specific, and well-centered.
4	Strong, with only limited omissions or drift.
3	Recognizable, but incomplete or materially mis-centered.
2	Substantial omissions, major factual errors, invented events, or a weak/generic response.
1	Mostly wrong, contradictory, hallucinated, or non-responsive.

The plot rubric applies these anchors consistently: a strong score requires the distinctive premise, central conflict, and dominant emphasis to be substantially preserved.

Output contract. Return one JSON object with `id`, `plotEvaluation`, `synopsisEvaluation`, and `keyMessageEvaluation`. Each task record contains typed, severity-tagged `errors`, task-specific 1–5 dimension scores and an `overall` score, and a concise `overallRationale`. No Markdown or commentary appears outside the schema; the complete field-level schema is retained in the experiment artifacts.

Cultural Prediction and Assessment Prompt

Identity. You are a professional film classification analyst. You infer audience suitability and country-specific motion-picture classifications from subtitles only.

Task. Given subtitle entries for one film, predict the cultural content labels in one English JSON object: age-suitability prediction and country-specific motion-picture rating prediction.

Input constraint. Use only the provided subtitles. The subtitles may be English or translated subtitles in another language, but every output field and rationale must be written in English. The user message specifies `inputLanguage`; treat that language label as authoritative. Do not rely on film title, cast, franchise knowledge, reviews, ratings databases, trailers, images, or memorized outside knowledge. If a content issue is not inferable from subtitles, do not invent it.

Classification principles. Treat ratings as audience-suitability classifications. Consider subtitle-visible evidence such as profanity, threats, fear, violence described in dialogue, sexual references, substance references, mature themes, emotional intensity, crime, death, and disturbing situations. Be conservative when visual evidence would be required but is absent from subtitles.
Country ratings are culturally situated. Predict each country independently using that country's allowed label set and ordered scale. Do not collapse countries into a universal rating.

Allowed labels. Age suitability: 2+, 3+, 4+, 5+, 6+, 7+, 8+, 9+, 10+, 11+, 12+, 13+, 14+, 15+, 16+, 17+, 18+.

Country-specific motion-picture labels:

Country	Allowed labels
Australia	G, PG, M, MA15+, R18+
Brazil	Livre, 10, 12, 14, 16, 18
France	Tous publics, 12, 16, 18
Germany	0, 6, 12, 16, 18
Netherlands	AL, 6, 9, 12, 14, 16, 18
Singapore	G, PG, PG13, NC16, M18, R21
South Korea	All, 12, 15, 19
Sweden	Btl, 7, 11, 15
United Kingdom	U, PG, 12, 15, 18
United States	G, PG, PG-13, R, NC-17

Output rules. Return only valid JSON that conforms to the provided schema. Predict exactly one age-suitability label and exactly one rating for each of the ten countries. Use concise rationales that reference subtitle-visible evidence generally, not raw offensive wording. Do not include Markdown, commentary, confidence scores, wrapper keys, or snake-case aliases.

Abbreviated schema illustration: the actual response includes one country object for each of the ten countries.

```
{
  "id": 0,
  "ageSuitabilityPrediction": {
    "ageRating": "13+",
    "rationale": "Concise subtitle-visible rationale."
  },
  "countryMotionPictureRatingPrediction": [
    {
      "country": "Australia",
      "rating": "M",
      "rationale": "Concise subtitle-visible rationale."
    }
  ]
}
```

B.4 Independent Narrative-Judge Agreement

We assess agreement between Claude Sonnet 5[1] and GPT-5.6 Luna on the same model-generated narrative outputs using shared references and the same two-stage evaluation rubric. The audit samples 50 films and nine evaluated models, yielding 450 model–film cases. Claude Sonnet 5 uses its default settings. Complete plot, synopsis, and key-message scores are available for every case, giving 1,350 paired task ratings.

Table 17 shows high agreement across the pooled ratings. Pearson correlation is $r = 0.808$, Spearman correlation is $\rho = 0.815$, and quadratic weighted $\kappa = 0.778$. Film-clustered bootstrap 95% confidence intervals are $[0.780, 0.833]$, $[0.785, 0.839]$, and $[0.749, 0.803]$, respectively. Exact agreement is 56.8%, while 98.4% of paired ratings differ by at most one point. Claude Sonnet 5 assigns scores 0.19 points lower on average than GPT-5.6 Luna (95% CI $[-0.253, -0.122]$). Thus, the two judges show closely aligned relative scoring and near-universal agreement within one point, alongside a small systematic difference in score level.

Table 17: Agreement between Claude Sonnet 5 and GPT-5.6 Luna on narrative-task scores in the 50-film audit. The analysis covers all 450 model–film cases from nine evaluated models, yielding 1,350 paired task ratings across plot, synopsis, and key message.

Task	Paired scores	Pearson r	Spearman ρ	Weighted κ	Exact (%)	Within 1 (%)	Mean Δ
All three tasks	1350	0.808	0.815	0.778	56.8	98.4	-0.192
Plot	450	0.761	0.710	0.675	43.4	96.9	-0.438
Synopsis	450	0.750	0.751	0.729	61.5	98.2	-0.084
Key message	450	0.766	0.776	0.763	65.5	100.0	-0.053

Note. Each task is scored on a 1–5 scale. Mean Δ is Claude Sonnet 5 minus GPT-5.6 Luna, so negative values indicate lower Claude Sonnet 5 ratings. Weighted κ denotes quadratic weighted κ. For the pooled Pearson correlation, Spearman correlation, weighted κ, and mean difference, 95% confidence intervals use 5,000 film-clustered bootstrap resamples. Task-specific rows report point estimates.

B.5 Response-Format Reliability

A response is scored only after it satisfies the task schema. If an identifiable prediction is wrapped in Markdown or contains a strictly syntactic JSON defect, deterministic repair may correct the serialization before validation is repeated. A local model may first receive one formatting-only retry with the same evidence. Neither operation supplies a gold label or changes generated text, predicted labels, ratings, categories, or evidence indices.

Responses with no usable prediction, provider errors, or substantial truncation remain failures. Table 18 and Figure 9 isolate this reliability layer from semantic task accuracy.

Table 18: Structured-output recovery audit. AR, FA, VI, and EN denote Arabic, Persian, Vietnamese, and English, respectively. Arrows indicate the preferred direction: failure rates are lower-is-better (↓), while successful recovery is higher-is-better (↑). Percentages summarize serialization and provider-cleanup outcomes for saved raw model outputs. Recovery is restricted to structural parsing and schema validation; semantic labels, predictions, evidence, and scores are never altered.

Model	Task	Language	Initial structural failure (%) ↓	Structurally recovered (%) ↑	Remaining provider failure (%) ↓
Qwen3 235B	Narrative	AR	7.0	7.0	0.0
Qwen3 235B	Narrative	FA	7.5	7.5	0.0
Qwen3 235B	Narrative	VI	8.5	8.5	0.0
Llama 4 Scout 17B	Language-Safety	EN	1.5	1.5	0.0

B.6 Title/Year Recall and Character-Name Counterfactuals

We conduct two targeted probes to separate information recoverable from film identity cues from information used during subtitle-conditioned generation. The first measures how much task performance can be recovered from title and release year alone. The second measures how narrative

[1] https://platform.claude.com/docs/en/models/sonnet-5/

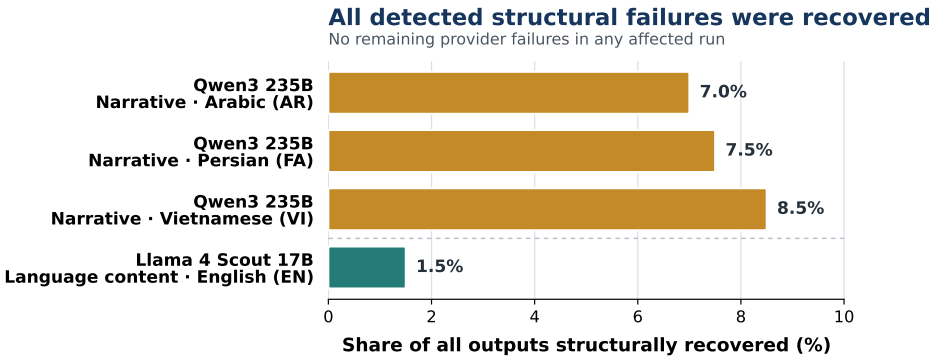


Figure 9: Response-format recovery in affected runs. Repairs modify serialization only; they do not change task content or scores.

generation responds when central-character names are systematically altered while the surrounding subtitle sequence and corresponding references are preserved.

Experimental conditions. The standard benchmark provides subtitle text, the input-language identifier, and task instructions/schema, but not film titles, release years, cast, reference narratives, or gold labels. For the title/year probe, we select 50 films evaluated with English subtitles evenly across five strata of narrative-judge difficulty. Each model receives only the film title and release year and generates a plot, synopsis, key message, and genre labels. We compare these outputs with the standard subtitle-only outputs for the same 50 films.

For the character-name counterfactual, we select 20 films, four from each difficulty stratum, and exchange two manually screened central-character aliases throughout each film's English subtitles and narrative references. Models then generate from the transformed subtitles without access to the film title or release year. We compare these generations with the standard subtitle-only outputs for the same films and score them against the correspondingly transformed references. Genre labels are unchanged by the transformation and are therefore excluded from the paired counterfactual narrative comparison.

Scoring and uncertainty. GPT-5.6 Luna applies the same two-stage reference-based rubric used for the main narrative evaluation. Existing subtitle-only scores are reused, and all 350 new outputs are collected successfully: 50 title/year and 20 counterfactual generations for each of five models. Tables 19 and 20 report per-task means and paired changes in the mean of plot, synopsis, and key-message scores on the 1–5 scale. Uncertainty is estimated using paired film-level 95% bootstrap intervals. Table 21 reports multi-label genre micro-F1 and exact match for the paired 50-film subtitle-only and title/year conditions.

Table 19: Narrative performance with title and year only versus subtitles. Task entries are mean rubric scores on a 1–5 scale, shown as subtitle-only/title-year-only for paired films. The final column is the paired change in the per-film mean across the three tasks, with a 95% film-level bootstrap interval.

Model	Plot (S/T)	Synopsis (S/T)	Key message (S/T)	Mean change [95% CI]
GPT-5.6 Sol	4.58 / 4.68	3.38 / 3.54	3.00 / 2.96	$+0.07$ [-0.05, $+0.19$]
Gemini 3.8 Flash	4.58 / 4.68	3.58 / 3.36	3.14 / 3.20	-0.02 [-0.13, $+0.09$]
Claude Haiku 4.5	4.12 / 3.94	2.48 / 2.24	2.94 / 2.70	-0.22 [-0.40, -0.05]
DeepSeek V4 Pro	4.38 / 4.52	3.30 / 2.96	2.98 / 3.06	-0.04 [-0.20, $+0.13$]
Llama 4 Scout	2.82 / 2.60	1.88 / 1.60	2.50 / 2.30	-0.23 [-0.47, -0.01]

Note. S/T denotes subtitle-only/title-year-only. $n = 50$ paired films for every model. A positive change favors title/year-only generation; a negative change favors subtitle-conditioned generation.

Table 20: Narrative performance on the character-name counterfactual subset. Task entries are mean rubric scores on a 1–5 scale, shown as subtitle-only/counterfactual-subtitle for paired films. The final column is the paired change in the per-film mean across the three tasks, with a 95% film-level bootstrap interval.

Model	Plot (S/C)	Synopsis (S/C)	Key message (S/C)	Mean change [95% CI]
GPT-5.6 Sol	4.70 / 3.80	3.30 / 2.70	3.00 / 3.05	-0.48 [-0.82, -0.17]
Gemini 3.8 Flash	4.65 / 4.50	3.55 / 2.80	3.15 / 3.00	-0.35 [-0.57, -0.15]
Claude Haiku 4.5	4.10 / 3.95	2.40 / 2.35	2.90 / 2.75	-0.12 [-0.30, $+0.07$]
DeepSeek V4 Pro	4.45 / 4.25	3.20 / 2.60	2.95 / 3.15	-0.20 [-0.45, $+0.03$]
Llama 4 Scout	2.80 / 2.65	1.90 / 1.70	2.55 / 2.25	-0.22 [-0.53, $+0.10$]

Note. S/C denotes subtitle-only/counterfactual-subtitle. $n = 20$ paired films (four sampled from each of five narrative-difficulty strata). In the counterfactual condition, two character aliases are exchanged in the subtitles and narrative references; the film title and year are withheld. Negative changes indicate lower rubric scores under the counterfactual condition.

Table 21: Genre prediction from subtitles versus film title and year on the paired 50-film sample. Micro-F1 summarizes multi-label decisions across all genre labels; exact match requires the complete predicted genre set to match the reference.

Model	Micro-F1 (S)	Micro-F1 (T)	Change (pp)	Exact (S, %)	Exact (T, %)
GPT-5.6 Sol	75.7	82.3	$+6.6$	20.0	30.0
Gemini 3.8 Flash	79.8	84.7	$+4.9$	24.0	40.0
Claude Haiku 4.5	67.3	72.4	$+5.1$	14.0	8.0
DeepSeek V4 Pro	79.3	81.8	$+2.5$	30.0	34.0
Llama 4 Scout	61.2	72.2	$+11.0$	2.0	18.0

Note. S/T denotes subtitle-only/title-year-only; both conditions use the same 50 films and references. Micro-F1 values and changes are percentages and percentage points, respectively. Higher is better.

Film-identity cues carry substantial task signal. Title/year-only narrative performance remains close to the subtitle-only baseline for several models: paired mean changes range from -0.23 to $+0.07$. The intervals exclude zero for Claude Haiku 4.5 and Llama 4 Scout, both favoring subtitle-conditioned generation. Genre prediction shows a stronger identity-cue effect: all five models achieve higher micro-F1 from title and year alone, with gains of 2.5–11.0 percentage points. This contrast shows that film identity alone provides substantial information for genre recognition, while detailed narrative reconstruction benefits more unevenly from subtitle evidence across models.

Narrative generation is sensitive to character-name perturbations. The counterfactual transformation lowers the paired narrative mean for all five models. The largest decreases occur for GPT-5.6 Sol (-0.48) and Gemini 3.8 Flash (-0.35), whose bootstrap intervals exclude zero; intervals for the other three models include zero. The consistent direction across models shows that altering central-character names can materially change subtitle-conditioned narrative generation. Because a name substitution can simultaneously change lexical form, gender cues, cultural associations, frequency, tokenization, and character familiarity, the intervention identifies sensitivity to the substitution as a whole rather than to any single property of the names. Together, the two probes show that benchmark performance reflects both information available from film identity and the model's response to evidence supplied in the subtitle context.

C Additional Evidence for RQ 1: Model Capability Profiles

C.1 English and Cross-Lingual Headline Results

Tables 22 and 23 separate the benchmark's task families under English subtitle input and the equal-weight average of Arabic, Indonesian, Persian, Romanian, and Vietnamese. Reporting the original metric units keeps task-level differences visible rather than collapsing them into the cross-task composite.

Task strengths already diverge under English input. DeepSeek V4 Pro approaches the two leading frontier models on English genre micro-F1 (83.2%) and age accuracy within one year (82.5%), while its country-rating exact match is 58.5%, compared with 78.8% for Gemini 3.8 Flash. GPT-5.6 Sol and Gemini are nearly tied on English synopsis generation (3.42 and 3.43), yet their country exact-match rates differ by 11.6 percentage points. Narrative quality and cultural prediction therefore need not move together even under the same subtitle language.

Model access does not determine a uniform capability profile. Across the six-language averages, DeepSeek V4 Pro exceeds Claude Haiku 4.5 on narrative overall, genre micro-F1, age prediction, country-rating exact match, and strict language-safety evidence F1. Its gaps from GPT-5.6 Sol vary considerably by task, from near-equal genre and age performance to larger differences in narrative reconstruction, country ratings, and evidence grounding.

Task-specific differences persist under non-English subtitle input. Gemini 3.8 Flash retains the highest genre micro-F1, age exact match, and country-rating exact match in both the English and five-language-average results. GPT-5.6 Sol achieves the highest cross-lingual synopsis score at 3.39/5, while the highest cross-lingual age exact-match rate is 33.8% for Gemini. Language-safety results are not included in this comparison because the evidence-grounding task is evaluated over English subtitle indices.

Table 22: English-only results. Arrows indicate the preferred direction: ↑ is higher-is-better and ↓ is lower-is-better. Narrative metrics are reference-based scores on a 1–5 scale; genre, age-threshold, and country exact-match metrics are percentages, while MAE metrics report error. Within each metric column, shading is normalized across the displayed models in the preferred direction: muted blue denotes narrative metrics and muted bronze denotes cultural metrics. Shade intensity is therefore not comparable across columns. Best values are bold, and second-best distinct values are underlined at the displayed precision; ties share the same emphasis.

Model	Family	Narrative Understanding and Generation					Cultural Prediction and Assessment				
		Plot ↑	Synopsis ↑	Key message ↑	Overall ↑	Genre micro-F1 ↑	Age exact ↑	Age ±1 ↑	Age MAE ↓	Country EM ↑	Country MAE ↓
GPT-5.6 Sol	Closed frontier	**4.17**	3.42	3.10	3.56	82.2	33.5	82.0	0.88	67.2	0.42
Gemini 3.8 Flash	Closed frontier	**4.17**	**3.43**	**3.23**	**3.61**	**84.7**	**36.5**	**83.0**	**0.85**	**78.8**	**0.29**
Claude Haiku 4.5	Closed frontier	3.62	2.54	2.92	3.02	72.2	29.5	78.0	0.99	58.0	0.53
GPT-5 Nano	Closed baseline	3.21	2.06	2.71	2.66	74.8	19.5	60.5	1.55	43.8	0.75
Gemini 3.5 Flash Lite	Closed baseline	3.97	3.02	3.06	3.35	81.8	32.5	75.5	0.99	66.3	0.44
DeepSeek V4 Pro 1.6T	Open-weight	3.91	3.08	2.90	3.30	83.2	34.5	82.5	0.88	58.5	0.51
Qwen3 235B	Open-weight	3.43	2.46	2.88	2.92	69.9	21.5	63.5	1.36	51.7	0.60
Mistral 4 119B	Open-weight	2.98	2.12	2.52	2.54	70.0	17.5	56.0	1.77	49.4	0.65
Llama 4 Scout 17B	Open-weight	2.50	1.89	2.41	2.26	68.3	16.5	44.0	2.21	33.1	0.92

C.2 Paired Comparisons of Leading Composite Scores

We assess uncertainty in the composite-score gaps among the three leading models. Gemini 3.8 Flash, the highest-scoring model, is compared with GPT-5.6 Sol and with DeepSeek V4 Pro 1.6T, the highest-scoring open-weight model. For each comparison, films are resampled as paired clusters while retaining all subtitle-language conditions and the English-only language-content component. The composite score is then recomputed from its four components. We report percentile 95% confidence intervals from 10,000 paired film-level bootstrap resamples.

We additionally conduct two-sided paired randomization tests by swapping the two models' complete outputs within each film cluster and recomputing the composite difference over 9,999 random swaps. Holm correction is applied across the two planned comparisons.

Gemini's composite score exceeds both comparators. Using comparator minus Gemini as the paired difference, GPT-5.6 Sol differs by -3.34 points (95% CI $[-3.86, -2.84]$), while DeepSeek V4 Pro 1.6T differs by -12.52 points (95% CI $[-13.61, -11.44]$). Both intervals lie below zero, and both Holm-adjusted randomization-test p-values are < 0.001 (Table 25). These estimates characterize

Table 23: Cross-lingual results averaged over Arabic, Indonesian, Persian, Romanian, and Vietnamese subtitle inputs; English is excluded. Arrows indicate the preferred direction: ↑ is higher-is-better and ↓ is lower-is-better. Narrative metrics are reference-based scores on a 1–5 scale; genre, age-threshold, and country exact-match metrics are percentages, while MAE metrics report error. Within each metric column, shading is normalized across the displayed models in the preferred direction: muted blue denotes narrative metrics and muted bronze denotes cultural metrics. Shade intensity is therefore not comparable across columns. Best values are bold, and second-best distinct values are underlined at the displayed precision; ties share the same emphasis.

Model	Family	**Narrative Understanding and Generation**					**Cultural Prediction and Assessment**				
		Plot ↑	Synopsis ↑	Key message ↑	Overall ↑	Genre micro-F1 ↑	Age exact ↑	Age ±1 ↑	Age MAE ↓	Country EM ↑	Country MAE ↓
GPT-5.6 Sol	Closed frontier	4.08	**3.39**	**3.08**	**3.52**	81.4	30.6	78.1	1.01	65.1	0.45
Gemini 3.8 Flash	Closed frontier	**4.16**	3.32	**3.08**	**3.52**	**85.0**	**33.8**	**81.1**	**0.95**	**77.7**	**0.31**
Claude Haiku 4.5	Closed frontier	3.42	2.42	2.82	2.89	71.3	27.0	71.7	1.19	53.6	0.58
GPT-5 Nano	Closed baseline	2.92	1.87	2.62	2.47	72.0	20.9	60.8	1.56	43.9	0.77
Gemini 3.5 Flash Lite	Closed baseline	3.81	2.89	2.98	3.23	80.7	31.7	71.5	1.11	64.3	0.47
DeepSeek V4 Pro 1.6T	Open-weight	3.91	2.90	2.90	3.24	81.3	29.9	78.7	1.01	57.9	0.53
Qwen3 235B	Open-weight	3.34	2.27	2.73	2.78	68.0	22.1	58.4	1.49	47.1	0.65
Mistral 4 119B	Open-weight	2.77	1.97	2.52	2.42	66.2	19.8	54.2	1.78	47.0	0.69
Llama 4 Scout 17B	Open-weight	2.15	1.63	2.16	1.98	66.6	19.0	51.5	1.87	34.3	0.93

Table 24: Narrative generation scores averaged across all six subtitle languages. All models are evaluated with the same reference-based 1–5 rubric; higher is better (↑). Within each metric column, blue shading is normalized across the displayed models. Best values are bold, and second-best distinct values are underlined at the displayed precision; ties share the same emphasis.

Model	Plot ↑	Synopsis ↑	Key message ↑	Overall ↑
GPT-5.6 Sol	4.10	**3.40**	3.08	3.52
Gemini 3.8 Flash	**4.16**	3.34	**3.11**	**3.54**
Claude Haiku 4.5	3.45	2.44	2.84	2.91
GPT-5 Nano	2.97	1.90	2.64	2.50
Gemini 3.5 Flash Lite	3.83	2.91	2.99	3.25
DeepSeek V4 Pro 1.6T	3.91	2.93	2.90	3.25
Qwen3 235B	3.36	2.30	2.76	2.80
Mistral 4 119B	2.80	2.00	2.52	2.44
Llama 4 Scout 17B	2.21	1.67	2.20	2.03

Table 25: Paired film-level comparisons of the leading composite scores. Differences are computed as comparator minus Gemini 3.8 Flash, so negative values indicate a higher Gemini score.

Model	Composite score	Difference from Gemini (95% CI)	Holm-adjusted p
Gemini 3.8 Flash	78.04	Reference	—
GPT-5.6 Sol	74.70	-3.34 $[-3.86, -2.84]$	< 0.001
DeepSeek V4 Pro 1.6T	65.52	-12.52 $[-13.61, -11.44]$	< 0.001

Note. Confidence intervals are percentile intervals from 10,000 paired film-level bootstrap resamples. Two-sided paired randomization tests use 9,999 random within-film swaps of the compared models' complete outputs, with p-values Holm-adjusted across the two planned comparisons. The composite is recomputed from its four components in every resample.

film-level sampling uncertainty in the present evaluation and do not include run-to-run generation variability.

C.3 Narrative Failure Patterns

The structured evaluator assigns task-specific error tags before scoring every judged narrative output. We aggregate those tags here to reveal recurring failure mechanisms; they are evaluator diagnostics rather than independent human annotations.

Synopsis failures combine omission and invention. The evaluator marks a missing major event in 79.7% and an invented event in 71.1% of judged synopses. Plot errors instead most often concern wrong emphasis (54.7%), whereas key-message errors most often concern semantic mismatch (80.3%). These are output-level diagnostic incidences, not independent human labels.

Table 26: Narrative error patterns identified by the reference-based evaluator across the comparable model suite. *Share of errors* is the proportion of all error annotations within a narrative task assigned to each error type, while *outputs with error* is the percentage of judged outputs containing that error at least once. The final column reports the mean number of annotations of that type per output.

Task	Error type	Share of errors (%)	Outputs with error (%)	Avg. annotations/output
Plot	Wrong Emphasis	27.0	54.7	0.55
	Wrong Setup	18.5	37.5	0.38
	Missing Central Conflict	18.2	36.9	0.37
	Invented Event	13.4	27.0	0.27
	Too Vague	10.3	20.9	0.21
	Missing Protagonist Or Group	9.1	18.5	0.19
	Wrong Relationship	3.3	6.7	0.07
	Style Or Length Issue	0.1	0.3	0.00
Synopsis	Missing Major Event	23.4	79.7	0.82
	Invented Event	22.9	71.1	0.80
	Wrong Resolution	13.0	44.6	0.46
	Wrong Relationship	10.0	33.5	0.35
	Weak Causality	9.1	31.8	0.32
	Missing Resolution	8.0	27.9	0.28
	Missing Character Motivation	7.1	24.6	0.25
	Wrong Event Order	3.0	10.4	0.11
	Too Vague	2.2	7.7	0.08
	Style Or Length Issue	0.8	2.6	0.03
	Too Detailed	0.5	1.6	0.02
Key Message	Semantic Mismatch	44.1	80.3	0.80
	Too Generic	31.9	58.2	0.58
	Unsupported Theme	19.9	36.3	0.36
	Misses Positive Message	3.4	6.3	0.06
	Contradicted By Reference	0.3	0.5	0.00
	Style Or Length Issue	0.2	0.4	0.00
	Too Plot Like	0.2	0.4	0.00

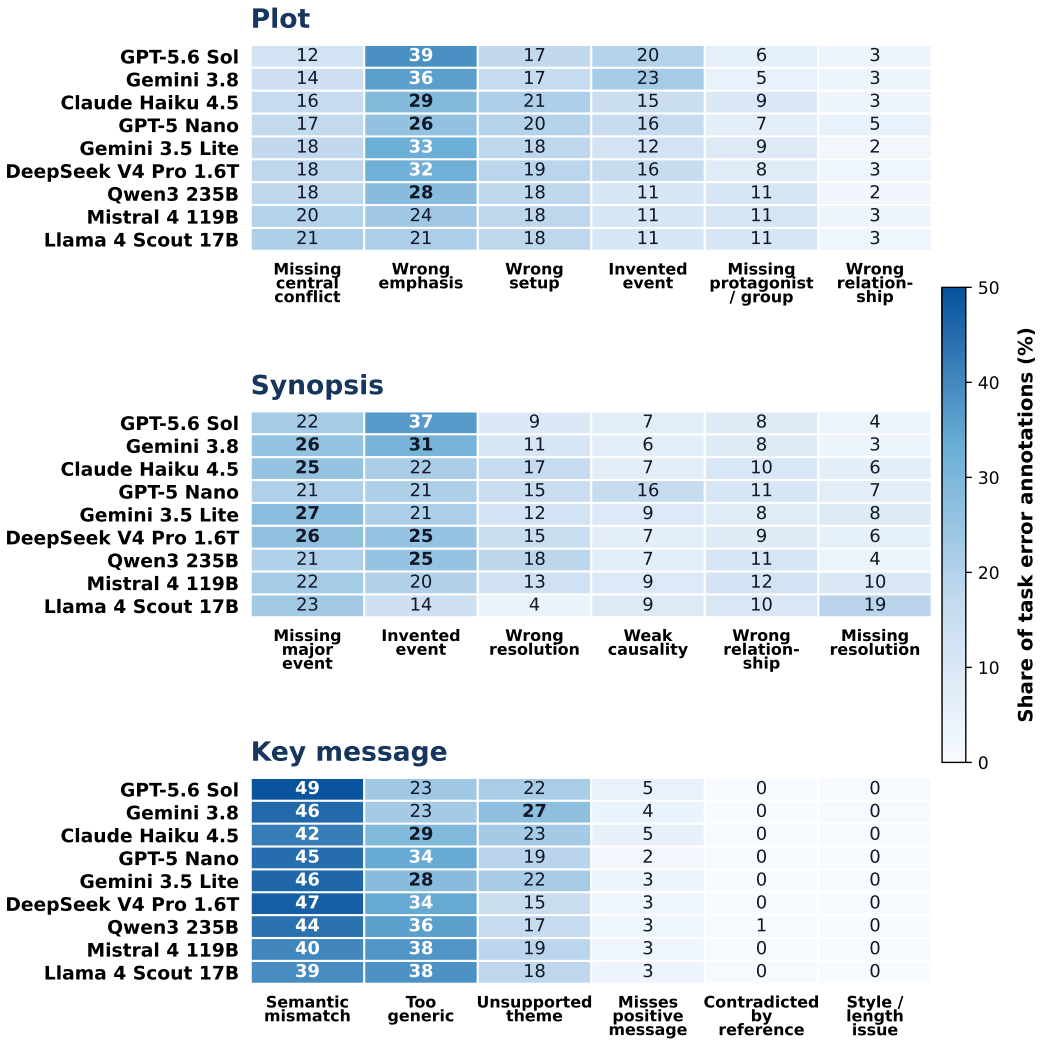


Figure 10: Narrative error profiles by model and subtask. Each cell reports the share of that subtask's structured error annotations assigned to the indicated type. Synopsis errors concentrate in missing and invented events; key-message errors concentrate in semantic mismatch and over-generic themes.

D Additional Evidence for RQ 2: Cross-Lingual Behavior

D.1 English-to-Cross-Lingual Overview

Table 27 compares English input with the equal-weight average of the five non-English subtitle settings on the same films. Narrative overall decreases for all nine models, although the magnitude varies considerably. Other tasks are less uniform: some model-language combinations improve on genre, age suitability, or country ratings even when narrative performance declines. The aggregate view therefore motivates the language-specific analyses that follow.

Table 27: English-to-cross-lingual comparison across headline tasks. EN denotes English, while CL is the equal-weight mean over Arabic, Indonesian, Persian, Romanian, and Vietnamese; Δ = CL − EN. Positive and negative markers indicate the direction of the observed change. Paired film-level uncertainty intervals are reported in Table 37.

Model	Family	Narrative Understanding and Generation						Cultural Prediction and Assessment					
		Overall (1–5)			Genre micro-F1 (%)			Age exact (%)			Country EM (%)		
		EN	CL	Δ	EN	CL	Δ	EN	CL	Δ	EN	CL	Δ
GPT-5.6 Sol	Closed frontier	3.56	3.52	↓-0.05	82.2	81.4	↓-0.8	33.5	30.6	↓-2.9	67.2	65.1	↓-2.1
Gemini 3.8 Flash	Closed frontier	3.61	3.52	↓-0.09	84.7	85.0	↑+0.3	36.5	33.8	↓-2.7	78.8	77.7	↓-1.1
Claude Haiku 4.5	Closed frontier	3.02	2.89	↓-0.14	72.2	71.3	↓-0.9	29.5	27.0	↓-2.5	58.0	53.6	↓-4.4
GPT-5 Nano	Closed baseline	2.66	2.47	↓-0.19	74.8	72.0	↓-2.8	19.5	20.9	↑+1.4	43.8	43.9	↑+0.1
Gemini 3.5 Flash Lite	Closed baseline	3.35	3.23	↓-0.13	81.8	80.7	↓-1.1	32.5	31.7	↓-0.8	66.3	64.3	↓-2.0
DeepSeek V4 Pro 1.6T	Open-weight/API	3.3	3.24	↓-0.06	83.2	81.3	↓-1.9	34.5	29.9	↓-4.6	58.5	57.9	↓-0.6
Qwen3 235B	Open-weight/API	2.92	2.78	↓-0.14	69.9	68.0	↓-1.9	21.5	22.1	↑+0.6	51.7	47.1	↓-4.6
Mistral 4 119B	Open-weight/API	2.54	2.42	↓-0.12	70.0	66.2	↓-3.8	17.5	19.8	↑+2.3	49.4	47.0	↓-2.4
Llama 4 Scout 17B	Open-weight/API	2.26	1.98	↓-0.28	68.3	66.6	↓-1.7	16.5	19.0	↑+2.5	33.1	34.3	↑+1.2

D.2 Narrative Scores by Subtitle Language

Plot recovery remains more reliable than event-complete reconstruction. The tables separate plot, synopsis, key-message, and their overall score for every model and subtitle language. Across models, the English means are 3.55 for plot, 2.67 for synopsis, and 2.86 for key message; under Persian input they are 3.20, 2.40, and 2.66. The plot–synopsis separation therefore persists within languages rather than arising only from the English setting.

Table 28: Plot-generation scores by model and subtitle-input language. Values use the same reference-based 1–5 rubric across models and languages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted blue shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	**4.17**	4.13	4.10	3.97	4.05	**4.17**
Gemini 3.8 Flash	**4.17**	**4.23**	**4.13**	**4.08**	**4.17**	**4.17**
Claude Haiku 4.5	3.62	3.38	3.53	3.13	3.48	3.58
GPT-5 Nano	3.21	2.93	3.03	2.64	2.98	3.02
Gemini 3.5 Flash Lite	3.97	3.83	3.88	3.69	3.79	3.85
DeepSeek V4 Pro 1.6T	3.91	3.96	3.94	3.75	3.88	4.03
Qwen3 235B	3.43	3.38	3.39	3.12	3.38	3.44
Mistral 4 119B	2.98	2.65	2.81	2.56	2.92	2.90
Llama 4 Scout 17B	2.50	2.02	2.27	1.86	2.33	2.29

Table 29: Synopsis-generation scores by model and subtitle-input language. Values use the same reference-based 1–5 rubric across models and languages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted blue shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	3.42	**3.39**	**3.45**	**3.28**	**3.40**	**3.43**
Gemini 3.8 Flash	**3.43**	3.35	3.35	3.23	3.31	3.36
Claude Haiku 4.5	2.54	2.41	2.50	2.28	2.50	2.44
GPT-5 Nano	2.06	1.81	1.92	1.75	1.94	1.95
Gemini 3.5 Flash Lite	3.02	2.89	2.90	2.83	2.90	2.94
DeepSeek V4 Pro 1.6T	3.08	3.00	2.88	2.87	2.91	2.86
Qwen3 235B	2.46	2.27	2.33	2.10	2.29	2.37
Mistral 4 119B	2.12	1.90	2.04	1.83	2.05	2.04
Llama 4 Scout 17B	1.89	1.57	1.64	1.49	1.72	1.72

Table 30: Key-message generation scores by model and subtitle-input language. Values use the same reference-based 1–5 rubric across models and languages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted blue shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	3.10	**3.13**	**3.11**	2.99	3.06	3.08
Gemini 3.8 Flash	**3.23**	3.06	**3.11**	**3.02**	**3.10**	**3.12**
Claude Haiku 4.5	2.92	2.85	2.87	2.67	2.87	2.85
GPT-5 Nano	2.71	2.60	2.65	2.45	2.71	2.70
Gemini 3.5 Flash Lite	3.06	3.05	2.96	2.90	2.97	3.02
DeepSeek V4 Pro 1.6T	2.90	2.92	2.92	2.84	2.88	2.95
Qwen3 235B	2.88	2.73	2.76	2.62	2.75	2.79
Mistral 4 119B	2.52	2.42	2.60	2.40	2.58	2.62
Llama 4 Scout 17B	2.41	2.04	2.23	2.02	2.29	2.21

Table 31: Narrative-overall scores by model and subtitle-input language. Values use the same reference-based 1–5 rubric across models and languages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted blue shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	3.56	**3.55**	**3.55**	3.41	3.50	**3.56**
Gemini 3.8 Flash	**3.61**	**3.55**	3.53	**3.44**	**3.53**	3.55
Claude Haiku 4.5	3.02	2.88	2.96	2.69	2.95	2.96
GPT-5 Nano	2.66	2.45	2.53	2.28	2.54	2.56
Gemini 3.5 Flash Lite	3.35	3.26	3.25	3.14	3.22	3.27
DeepSeek V4 Pro 1.6T	3.30	3.29	3.25	3.15	3.22	3.28
Qwen3 235B	2.92	2.79	2.83	2.61	2.81	2.87
Mistral 4 119B	2.54	2.32	2.48	2.26	2.52	2.52
Llama 4 Scout 17B	2.26	1.88	2.04	1.79	2.11	2.07

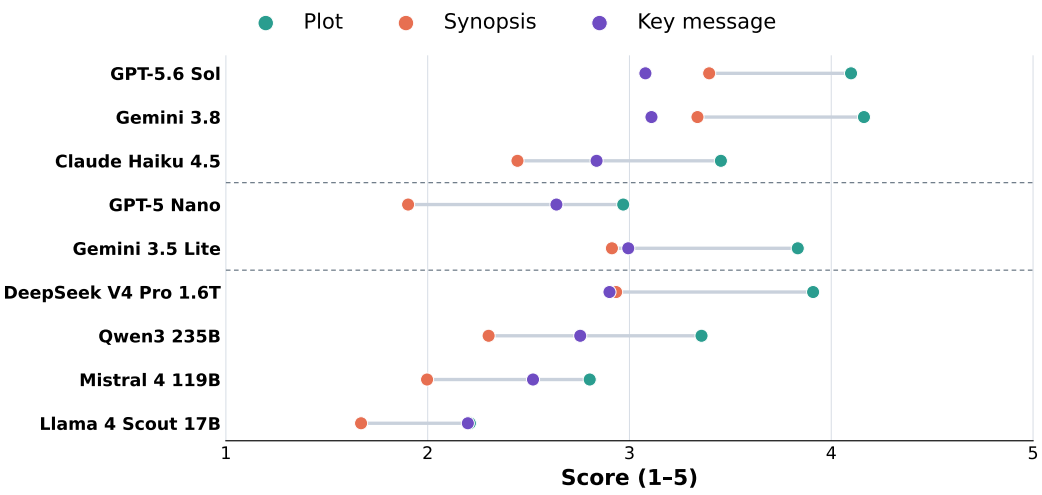


Figure 11: Six-language-average scores for plot, synopsis, and key-message generation. Each row represents one model, and the connecting segment highlights the separation between premise-level plot recovery and fuller synopsis reconstruction.

D.3 Genre Stability and Label-Specific Failures

Cross-lingual genre changes are model dependent. From English to the five-language average, Mistral 4 119B loses 3.8 percentage points of genre micro-F1, while Gemini 3.8 Flash changes by only $+0.3$ points; Gemini's paired interval includes zero. These differences are small for some models and more pronounced for others, reinforcing the need for model-specific cross-lingual analysis.

High overlap does not imply exact label-set recovery. Gemini 3.8 reaches 84.9% six-language micro-F1 but 40.8% exact match; DeepSeek reaches 81.6% and 34.0%. A model may therefore recognize several relevant genres while adding or omitting another.

Agreement is strongest on several recognizable genres. Sol and Gemini differ by only 0.7 points on Comedy and 0.4 on Animation, and both reach 100.0% on Sport and Documentary. Comedy is common (38.0% of evaluated films), making its agreement more persuasive than Sport (2.0%) or Documentary (0.5%). Film-Noir is omitted because the evaluated set contains no positive examples.

Genre errors are label-selective rather than uniform. DeepSeek nearly matches Gemini on Science Fiction (85.5% versus 85.9%) and Animation (90.7% versus 91.7%), and exceeds Sol on Biography (86.6% versus 82.5%). By contrast, Qwen3 and Mistral 4 score 5.1% and 6.8% on Biography after missing 111 and 110 of 114 positive film–language instances with no false positives: a systematic omission hidden by aggregate F1.

D.4 Paired Cross-Lingual Uncertainty

Table 37 resamples matched films across subtitle languages rather than treating translations as independent films. The small narrative declines observed for GPT-5.6 Sol and DeepSeek V4 Pro have intervals that include zero, as do Llama 4 Scout's positive country point estimate and several small positive age-exact changes. Llama's narrative decline remains clearly negative. The paired analysis separates stable directional changes from visually small point-estimate differences.

Table 32: Genre-prediction metrics averaged across all six subtitle languages; higher is better (↑) for every metric. Genre prediction is multi-label, and exact match requires the complete predicted label set to equal the reference set. Within each metric column, blue shading is normalized across the displayed models. Best values are bold, and second-best distinct values are underlined at the displayed precision; ties share the same emphasis.

Model	Exact (%) ↑	Jaccard (%) ↑	Micro-F1 (%) ↑	Macro-F1 (%) ↑
GPT-5.6 Sol	28.8	72.1	81.5	<u>80.5</u>
Gemini 3.8 Flash	**40.8**	**77.7**	**84.9**	**83.3**
Claude Haiku 4.5	16.9	60.4	71.5	69.8
GPT-5 Nano	11.5	59.5	72.4	70.4
Gemini 3.5 Flash Lite	30.7	71.9	80.9	77.8
DeepSeek V4 Pro 1.6T	<u>34.0</u>	<u>73.2</u>	<u>81.6</u>	<u>80.5</u>
Qwen3 235B	13.0	55.9	68.3	59.0
Mistral 4 119B	9.7	53.4	66.8	63.8
Llama 4 Scout 17B	11.2	54.4	66.9	61.7

Table 33: Genre micro-F1 by model and subtitle-input language. Values are percentages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted blue shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	82.2	<u>81.9</u>	81.7	<u>79.9</u>	<u>81.4</u>	82.2
Gemini 3.8 Flash	**84.7**	**85.0**	**85.5**	**84.4**	**84.3**	**85.7**
Claude Haiku 4.5	72.2	71.2	73.6	67.5	72.0	72.2
GPT-5 Nano	74.8	71.2	73.2	68.4	72.7	74.3
Gemini 3.5 Flash Lite	81.8	79.7	81.9	79.5	80.0	82.6
DeepSeek V4 Pro 1.6T	<u>83.2</u>	80.6	<u>82.8</u>	79.4	81.1	<u>82.7</u>
Qwen3 235B	69.9	66.8	69.0	65.4	69.1	69.5
Mistral 4 119B	70.0	64.1	67.9	63.7	68.3	67.0
Llama 4 Scout 17B	68.3	65.8	67.7	64.2	67.1	68.4

Table 34: Genre macro-F1 by model and subtitle-input language. Values are percentages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted blue shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	<u>80.2</u>	<u>80.9</u>	80.9	<u>79.4</u>	<u>81.2</u>	80.5
Gemini 3.8 Flash	80.0	**84.2**	**84.8**	**83.2**	**83.5**	**84.1**
Claude Haiku 4.5	71.3	68.0	72.7	65.5	70.5	70.7
GPT-5 Nano	73.3	68.9	70.6	66.0	71.1	72.3
Gemini 3.5 Flash Lite	77.9	77.0	79.6	75.7	78.1	78.8
DeepSeek V4 Pro 1.6T	**82.1**	79.7	<u>81.7</u>	77.9	80.8	<u>80.6</u>
Qwen3 235B	60.9	57.5	59.5	56.4	58.8	61.2
Mistral 4 119B	67.6	61.1	66.2	56.7	67.0	64.3
Llama 4 Scout 17B	62.5	60.9	61.8	60.3	63.0	61.9

Table 35: Genre exact match by model and subtitle-input language. Values are percentages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted blue shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	28.0	29.0	29.0	27.5	30.0	29.5
Gemini 3.8 Flash	**37.0**	**40.0**	**40.5**	**42.5**	**42.0**	**43.0**
Claude Haiku 4.5	17.5	16.0	20.0	14.0	17.5	16.5
GPT-5 Nano	13.5	9.0	12.5	9.0	13.0	12.0
Gemini 3.5 Flash Lite	31.0	28.0	32.5	29.5	29.5	33.5
DeepSeek V4 Pro 1.6T	**37.0**	32.0	35.0	31.0	32.5	36.5
Qwen3 235B	16.0	10.0	14.0	11.0	14.5	12.5
Mistral 4 119B	11.5	8.0	12.0	6.5	9.5	10.5
Llama 4 Scout 17B	12.5	10.5	11.0	8.5	12.0	12.5

Table 36: Genre-wise F1 scores pooled across the six subtitle-input languages for each model. Values are percentages. *Gold* reports label prevalence in the evaluated film set, providing context for less frequent genres. Each genre is evaluated as a one-versus-rest binary classification problem; labels with no positive gold examples are omitted because their F1 is not informative. Within each genre row, blue shading compares performance across models. Best values are bold, and second-best distinct values are underlined at the displayed precision. Model headings link to the corresponding provider or model pages.

Genre	Gold (%)	Sol	Gemini 3.8	Claude	Nano	Gemini 3.5	DeepSeek	Qwen3	Mistral 4	Llama 4
Drama	50.0	83.0	**87.2**	79.5	73.5	84.9	86.1	76.8	74.5	71.9
Comedy	38.0	90.6	**91.3**	77.1	69.3	88.9	89.2	75.0	71.3	79.2
Adventure	36.5	82.3	**84.7**	55.6	80.0	83.3	78.3	68.6	63.9	77.7
Action	34.5	87.6	**90.4**	78.6	80.3	89.9	87.2	78.5	60.8	72.4
Thriller	30.0	78.8	**79.8**	73.4	73.8	77.2	78.2	67.3	67.7	68.3
Fantasy	22.0	72.1	**78.9**	65.9	68.0	67.6	70.7	61.1	62.9	63.5
Family	17.0	80.4	**82.9**	77.5	64.6	60.6	80.4	66.9	73.2	51.0
Sci-Fi	17.0	83.7	**85.9**	79.3	76.7	80.4	85.5	81.4	82.3	54.7
Mystery	16.5	71.0	**76.7**	21.9	63.4	69.3	69.0	22.4	57.2	62.1
Crime	16.0	79.1	**83.1**	74.0	70.3	82.5	79.1	68.0	63.8	52.5
Animation	13.0	91.3	**91.7**	62.9	83.4	91.0	90.7	78.2	72.4	39.6
Romance	13.0	67.5	**76.7**	69.8	70.6	75.6	75.7	68.3	56.9	66.5
Horror	12.5	**88.3**	87.9	76.7	73.2	82.3	83.7	52.8	66.4	46.3
Biography	9.5	82.5	**95.0**	68.6	82.3	93.2	86.6	5.1	6.8	65.1
Musical	6.5	59.5	**63.2**	20.7	63.0	35.8	44.0	37.1	25.0	11.8
History	5.5	69.2	67.2	**76.6**	55.6	67.6	69.8	42.2	45.8	37.0
Music	4.0	61.5	**74.4**	59.5	73.6	52.6	65.9	47.5	59.8	56.8
War	3.5	78.8	82.0	**89.1**	65.6	83.1	86.3	70.6	79.1	84.0
Sport	2.0	**100.0**	**100.0**	85.7	81.2	97.9	**100.0**	0.0	85.7	62.9
Western	2.0	**83.7**	**83.7**	75.0	69.8	73.7	**83.7**	76.9	80.0	80.0
Documentary	0.5	**100.0**	**100.0**	**100.0**	27.2	**100.0**	**100.0**	**100.0**	85.7	**100.0**

Table 37: Paired English-to-cross-lingual differences with percentile confidence intervals. Each Δ is computed on the same films as the mean over the five non-English subtitle-input languages minus English. Point estimates are calculated before rounding and displayed to at most two decimal places. Confidence intervals are obtained from 2,000 film-level bootstrap resamples that preserve the six-language pairing; genre micro-F1 is recomputed from pooled label decisions within each resample. Intervals containing zero do not indicate a consistently signed change.

Model	Narrative Δ	Genre F1 Δ (pp)	Age exact Δ (pp)	Country EM Δ (pp)
GPT-5.6 Sol	−0.05 [-0.11, +0.02]	−0.8 [-1.71, +0.30]	−2.9 [-7.30, +1.40]	−2.1 [-3.51, -0.72]
Gemini 3.8 Flash	−0.09 [-0.15, -0.02]	+0.3 [-0.62, +1.12]	−2.7 [-6.90, +1.30]	−1.1 [-2.10, -0.02]
Claude Haiku 4.5	−0.14 [-0.2, -0.07]	−0.9 [-2.27, +0.53]	−2.5 [-7.30, +2.00]	−4.4 [-6.56, -2.25]
GPT-5 Nano	−0.19 [-0.25, -0.14]	−2.8 [-4.51, -1.17]	+1.4 [-2.80, +5.50]	+0.1 [-2.58, +2.59]
Gemini 3.5 Flash Lite	−0.13 [-0.19, -0.06]	−1.1 [-2.24, +0.00]	−0.8 [-5.10, +3.40]	−2.0 [-3.53, -0.42]
DeepSeek V4 Pro 1.6T	−0.06 [-0.12, +0.01]	−1.9 [-3.01, -0.81]	−4.6 [-9.60, +0.00]	−0.6 [-2.33, +1.37]
Qwen3 235B	−0.14 [-0.2, -0.08]	−1.9 [-3.52, -0.25]	+0.6 [-4.20, +5.30]	−4.6 [-7.04, -2.26]
Mistral 4 119B	−0.12 [-0.18, -0.06]	−3.8 [-5.76, -1.83]	+2.3 [-2.10, +6.50]	−2.4 [-4.70, -0.01]
Llama 4 Scout 17B	−0.28 [-0.35, -0.22]	−1.7 [-3.45, +0.22]	+2.5 [-2.50, +7.30]	+1.2 [-1.70, +4.15]

E Additional Evidence for RQ 3: Age and National-Rating Calibration

E.1 Age Error Magnitude and Direction

Near-miss accuracy conceals asymmetric age behavior. The tables jointly report exact match, one- and two-year tolerance, MAE, and error direction. Across six subtitle languages, Gemini 3.8 Flash matches the reference age exactly in 34.2% of predictions and falls within one year in 81.4%; GPT-5.6 Sol reaches 31.1% and 78.8%, respectively. Their error directions differ sharply: Sol overpredicts age more often than it underpredicts (48.3% versus 20.6%), while Gemini shows the reverse pattern (15.2% versus 50.6%).

Large errors also separate models that look similar under near-match metrics. DeepSeek V4 Pro and Sol each have only 5.4% of predictions more than two years from the reference, compared with 17.0% for Qwen3, 24.9% for Mistral 4, and 29.1% for Llama 4 Scout. Models with similar MAE or within-one-year accuracy can therefore exhibit very different error magnitude and direction.

Table 38: Cultural prediction metrics averaged across all six subtitle languages. Arrows indicate the preferred direction: ↑ is higher-is-better and ↓ is lower-is-better. Age and country MAE are computed within their respective ordered label spaces. Within each metric column, bronze shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision; ties share the same emphasis.

Model	Age exact (%) ↑	Age ±1 (%) ↑	Age ±2 (%) ↑	Age MAE ↓	Country EM (%) ↑	Country MAE ↓
GPT-5.6 Sol	31.1	78.8	<u>94.6</u>	<u>0.98</u>	<u>65.4</u>	<u>0.44</u>
Gemini 3.8 Flash	**34.2**	**81.4**	**95.5**	**0.93**	**77.9**	**0.31**
Claude Haiku 4.5	27.4	72.8	90.4	1.16	54.3	0.57
GPT-5 Nano	20.7	60.8	80.4	1.56	43.9	0.76
Gemini 3.5 Flash Lite	<u>31.8</u>	72.2	91.0	1.09	64.7	0.46
DeepSeek V4 Pro 1.6T	30.7	<u>79.3</u>	<u>94.6</u>	<u>0.98</u>	58.0	0.52
Qwen3 235B	22.0	59.2	83.0	1.47	47.9	0.64
Mistral 4 119B	19.4	54.5	75.1	1.78	47.4	0.68
Llama 4 Scout 17B	18.6	50.2	70.9	1.93	34.1	0.93

E.2 Country-Level Calibration and Language Sensitivity

Raw accuracy and value above a national majority baseline can disagree. France has the lowest mean model exact match (31.6%) yet a 72.0% majority baseline; eight of nine models fall below that baseline, and 95.3% of nonzero model errors assign a stricter label. Gemini 3.8 Flash reaches 73.8%, only 1.8 points above the baseline. The United States has a lower 49.0% baseline but a

Table 39: Age exact match by model and subtitle-input language. Values are percentages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted bronze shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	33.5	31.5	31.5	26.0	**33.0**	31.0
Gemini 3.8 Flash	**36.5**	**34.5**	**33.0**	32.0	32.5	**37.0**
Claude Haiku 4.5	29.5	23.0	29.0	25.5	28.0	29.5
GPT-5 Nano	19.5	19.5	22.0	14.5	24.5	24.0
Gemini 3.5 Flash Lite	32.5	33.0	31.5	**33.0**	30.0	31.0
DeepSeek V4 Pro 1.6T	34.5	28.5	30.0	26.5	32.5	32.0
Qwen3 235B	21.5	20.5	21.5	27.5	22.0	19.0
Mistral 4 119B	17.5	20.0	19.0	21.5	18.5	20.0
Llama 4 Scout 17B	16.5	17.0	21.0	21.0	18.5	17.5

Table 40: Age accuracy within one year by model and subtitle-input language. Values are percentages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted bronze shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	82.0	78.5	77.5	77.5	78.0	79.0
Gemini 3.8 Flash	**83.0**	**81.5**	**82.5**	**79.5**	**82.0**	**80.0**
Claude Haiku 4.5	78.0	71.5	75.0	68.5	71.0	72.5
GPT-5 Nano	60.5	59.0	61.0	48.0	69.5	66.5
Gemini 3.5 Flash Lite	75.5	70.5	69.0	72.5	71.0	74.5
DeepSeek V4 Pro 1.6T	82.5	79.5	79.5	76.0	79.5	79.0
Qwen3 235B	63.5	57.5	57.0	59.0	57.5	61.0
Mistral 4 119B	56.0	55.5	54.5	52.0	52.0	57.0
Llama 4 Scout 17B	44.0	49.5	53.0	56.5	51.0	47.5

Table 41: Age accuracy within two years by model and subtitle-input language. Values are percentages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted bronze shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	**97.0**	94.0	94.0	94.0	**95.0**	93.5
Gemini 3.8 Flash	**97.0**	**96.0**	**94.5**	**95.0**	**95.0**	**95.5**
Claude Haiku 4.5	95.5	87.0	91.5	87.5	89.5	91.5
GPT-5 Nano	79.5	80.0	81.0	69.0	87.0	86.0
Gemini 3.5 Flash Lite	94.0	88.5	90.0	92.5	92.0	89.0
DeepSeek V4 Pro 1.6T	95.5	94.5	**94.5**	94.0	94.5	94.5
Qwen3 235B	86.5	84.5	81.5	82.0	82.5	81.0
Mistral 4 119B	75.0	78.0	75.0	73.5	71.0	78.0
Llama 4 Scout 17B	66.0	69.0	72.5	73.0	76.0	69.0

Table 42: Age mean absolute error by model and subtitle-input language. Values are reported in years; lower is better (↓). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted bronze shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	0.88	0.99	0.99	1.08	0.98	0.97
Gemini 3.8 Flash	**0.85**	**0.93**	**0.94**	**0.99**	**0.97**	**0.90**
Claude Haiku 4.5	0.99	1.25	1.08	1.34	1.18	1.09
GPT-5 Nano	1.55	1.60	1.54	2.02	1.33	1.33
Gemini 3.5 Flash Lite	0.99	1.14	1.14	1.09	1.12	1.07
DeepSeek V4 Pro 1.6T	0.88	1.02	0.98	1.09	0.98	0.95
Qwen3 235B	1.36	1.48	1.50	1.48	1.49	1.49
Mistral 4 119B	1.77	1.72	1.75	1.88	1.87	1.70
Llama 4 Scout 17B	2.21	1.95	1.78	1.83	1.84	1.95

Table 43: Age-prediction error magnitude and direction across all six subtitle-input languages. *Under* denotes predictions below the gold suitability age, while *over* denotes predictions above it. Exact, one-year, two-year, and larger-error shares partition all predictions; under- and overprediction rates exclude exact matches. Bronze shading and best/second-best emphasis apply only to Exact (higher is better) and > 2 years (lower is better); the remaining columns are descriptive.

Model	Exact (%)	Off by 1 (%)	Off by 2 (%)	¿2 years (%)	Under (%)	Over (%)
GPT-5.6 Sol	31.1	47.7	15.8	5.4	20.6	48.3
Gemini 3.8 Flash	**34.2**	47.2	14.1	**4.5**	50.6	15.2
Claude Haiku 4.5	27.4	45.3	17.7	9.6	31.9	40.7
GPT-5 Nano	20.7	40.1	19.7	19.6	19.1	60.2
Gemini 3.5 Flash Lite	31.8	40.3	18.8	9.0	50.3	17.8
DeepSeek V4 Pro 1.6T	30.7	48.7	15.2	5.4	29.8	39.6
Qwen3 235B	22.0	37.2	23.8	17.0	55.7	22.3
Mistral 4 119B	19.4	35.1	20.6	24.9	61.5	19.1
Llama 4 Scout 17B	18.6	31.7	20.7	29.1	48.3	33.1

41.8-point best-model lift. Country-level interpretation therefore requires both raw exact match and baseline-relative lift.

Country-rating performance does not follow access category. DeepSeek's 58.0% country exact match exceeds Claude's 54.3% overall and is higher in six of ten systems. It nevertheless trails Gemini in every country and Sol by 7.4 points overall. These reversals reinforce that model access category alone does not characterize cultural-prediction performance.

Sol exhibits different error regimes across national systems. Sol's nonzero errors are predominantly stricter in France (90.0%), Germany (95.3%), and the Netherlands (88.6%), but its lift over the respective majority baselines is -21.1, $+8.7$, and $+12.9$ points. Brazil is less concentrated: its majority baseline is 31.0%, and Sol improves it by 25.1 points. The same exact-match score can therefore arise from very different calibration regimes.

Subtitle-language effects are heterogeneous across models and countries. Relative to English, Romanian subtitles raise country-rating exact match by 3.7 points for Llama 4 Scout and 3.4 points for GPT-5 Nano. Averaged across models, Romanian input also improves Brazil by 2.3 points and Germany by 1.7 points, whereas all five non-English inputs lower United States accuracy. France is comparatively stable across subtitle languages, with mean model accuracy remaining between 30.0% and 32.5%. These patterns describe end-to-end language sensitivity rather than a change in the underlying gold rating.

Table 44: Country-wise motion-picture rating exact-match percentages by model, averaged across all six subtitle-input languages. Each national classification system retains its own ordered label space; higher is better within every column. Within each model column, bronze shading is normalized across countries to highlight relative country-level difficulty and is therefore not comparable across model columns. Best values are bold, and second-best distinct values are underlined at the displayed precision; ties share the same emphasis.

Country	Sol	Gemini 3.8	Claude Haiku	GPT-5 Nano	Gemini 3.5 Lite	DeepSeek V4 Pro 1.6T	Qwen3 235B	Mistral 4 119B	Llama 4 Scout 17B
United States	**87.1**	**90.8**	68.2	**72.0**	**81.9**	**73.9**	56.0	**63.3**	**53.3**
United Kingdom	77.0	89.2	60.5	53.3	76.0	68.7	57.7	56.8	44.8
South Korea	70.8	77.0	62.3	54.8	61.7	66.8	45.2	54.1	41.7
Australia	75.6	89.1	**71.6**	44.0	80.8	69.5	**69.2**	55.7	26.4
Singapore	60.2	64.4	44.2	39.1	52.8	51.0	36.4	35.2	24.8
Sweden	66.6	70.6	61.0	54.5	61.0	59.3	51.6	50.8	50.9
Germany	54.2	81.4	54.0	36.6	70.2	53.7	55.7	49.2	32.2
Brazil	56.1	68.6	46.3	28.8	59.8	52.6	46.4	40.1	28.1
Netherlands	55.4	74.0	54.7	35.4	64.8	52.4	41.1	45.9	31.5
France	50.9	73.8	20.3	20.5	37.4	32.2	19.6	22.2	7.4

Table 45: Country-level difficulty, improvement over baseline, and ordinal error direction. The majority baseline always predicts the most frequent gold label. Mean model EM averages the nine comparable models, while best EM is achieved by Gemini 3.8 Flash for every country and is averaged equally across the six subtitle-input languages. Lift is the difference between best EM and majority-baseline EM. Ordinal MAE averages model errors within each country's ordered label space. Stricter errors report the percentage of nonzero ordinal errors that assign a rating above the gold rating; exact predictions and mismatches between labels sharing the same ordinal value are excluded.

Country	Majority label	Baseline EM (%)	Mean model EM (%)	Best EM (%)	Lift (pp)	Ordinal MAE	Stricter errors (%)
Australia	M	42.5	64.7	89.1	+46.6	0.38	65.2
Brazil	14	31.0	47.4	68.6	+37.6	0.86	21.2
France	Tous publics	72.0	31.6	73.8	+1.8	0.90	95.3
Germany	12	45.5	54.1	81.4	+35.9	0.57	85.7
Netherlands	12	42.5	50.6	74.0	+31.5	0.89	70.4
Singapore	M18	26.0	45.4	64.4	+38.4	0.69	55.4
South Korea	15	36.0	59.4	77.0	+41.0	0.44	46.2
Sweden	15	40.0	58.5	70.6	+30.6	0.53	64.0
United Kingdom	15	47.0	64.9	89.2	+42.2	0.38	56.5
United States	R	49.0	71.8	90.8	+41.8	0.29	39.0

Table 46: Country-rating exact match by model and subtitle-input language. Values are percentages; higher is better (↑). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted bronze shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	67.2	65.8	65.7	62.4	65.8	65.6
Gemini 3.8 Flash	**78.8**	**78.6**	**77.5**	**76.5**	**77.2**	**78.6**
Claude Haiku 4.5	58.0	52.6	53.6	51.3	54.6	55.8
GPT-5 Nano	43.8	40.9	44.6	39.2	47.2	47.5
Gemini 3.5 Flash Lite	66.3	63.5	64.4	63.2	65.5	65.0
DeepSeek V4 Pro 1.6T	58.5	57.6	59.6	54.4	59.7	58.4
Qwen3 235B	51.7	46.9	48.4	44.9	47.4	48.0
Mistral 4 119B	49.4	45.1	47.9	44.9	47.1	49.8
Llama 4 Scout 17B	33.1	33.9	35.1	31.8	36.8	34.0

Table 47: Country-rating ordinal error by model and subtitle-input language. Values are measured in rating steps; lower is better (↓). EN, AR, ID, FA, RO, and VI denote English, Arabic, Indonesian, Persian, Romanian, and Vietnamese, respectively. Within each language column, muted bronze shading is normalized across the displayed models in the preferred direction. Best values are bold, and second-best distinct values are underlined at the displayed precision.

Model	EN	AR	ID	FA	RO	VI
GPT-5.6 Sol	0.42	0.44	0.43	0.48	0.45	0.43
Gemini 3.8 Flash	**0.29**	**0.30**	**0.31**	**0.33**	**0.32**	**0.30**
Claude Haiku 4.5	0.53	0.59	0.58	0.61	0.58	0.55
GPT-5 Nano	0.75	0.80	0.75	0.88	0.71	0.69
Gemini 3.5 Flash Lite	0.44	0.48	0.47	0.48	0.46	0.46
DeepSeek V4 Pro 1.6T	0.51	0.53	0.50	0.58	0.51	0.51
Qwen3 235B	0.60	0.64	0.64	0.68	0.66	0.64
Mistral 4 119B	0.65	0.72	0.66	0.73	0.69	0.64
Llama 4 Scout 17B	0.92	0.92	0.89	0.97	0.89	0.97

Table 48: Country-rating exact-match percentage by national classification system and subtitle-input language, averaged equally across the compared models. Columns correspond to English, Arabic, Indonesian, Persian, Romanian, and Vietnamese; higher is better (↑).

Country	English	Arabic	Indonesian	Persian	Romanian	Vietnamese
Australia	65.6	63.4	65.4	62.5	65.5	65.5
Brazil	47.7	45.1	48.4	43.9	49.9	49.4
France	31.6	31.2	32.5	29.9	32.5	31.9
Germany	54.2	53.4	56.3	50.6	55.9	54.5
Netherlands	50.8	50.1	51.5	46.6	52.6	51.9
Singapore	47.5	45.1	45.7	42.8	45.6	45.5
South Korea	61.9	59.6	58.3	57.6	59.6	59.3
Sweden	60.2	58.9	57.5	56.9	57.6	59.7
United Kingdom	67.9	62.0	64.8	61.3	65.9	67.3
United States	75.6	69.9	71.8	68.6	71.8	73.4

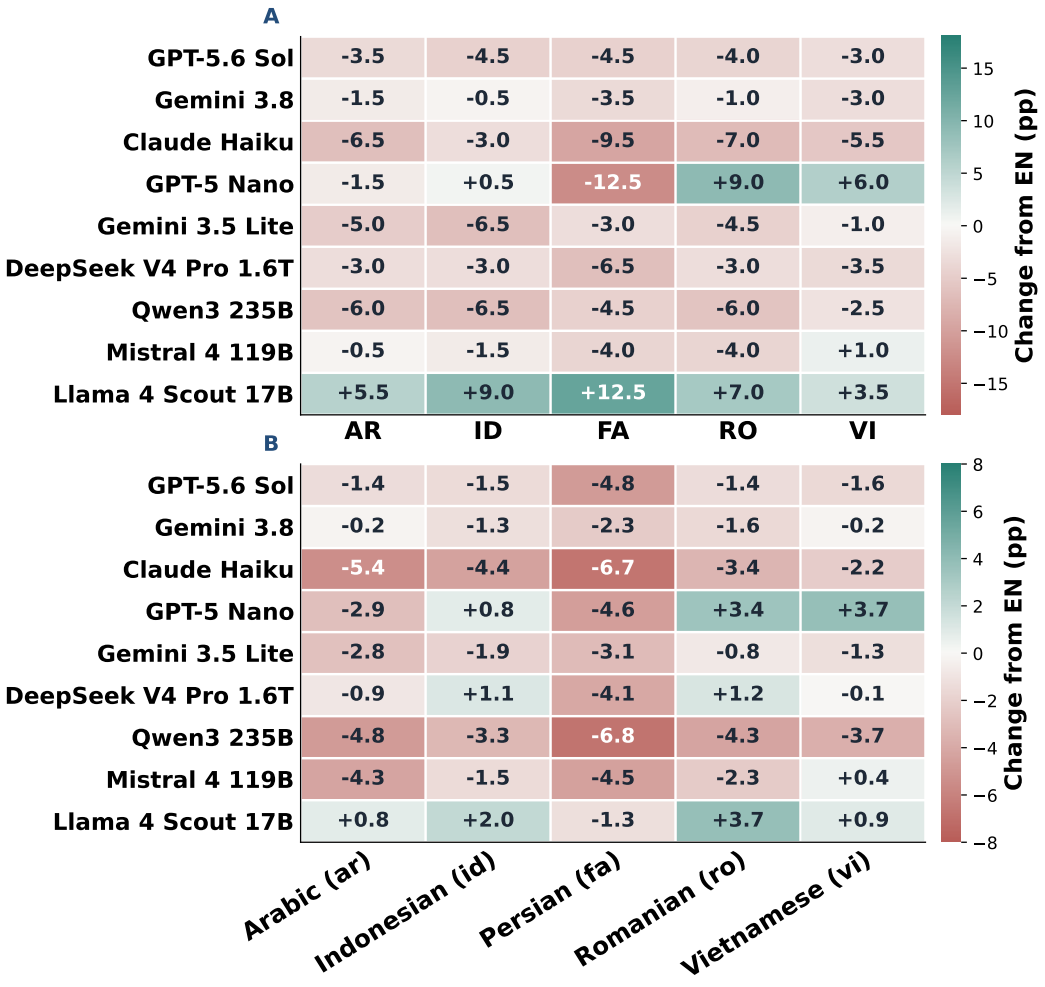


Figure 12: Change in cultural prediction relative to English subtitle input. The upper panel reports age accuracy within one year and the lower panel reports country-rating exact match, in percentage points, for each model under Arabic, Indonesian, Persian, Romanian, and Vietnamese subtitles. Absolute language-specific values are reported in the accompanying tables.

Table 49: Country-specific changes in exact-match accuracy under Arabic, Indonesian, Persian, Romanian, and Vietnamese subtitle inputs relative to English. Values are percentage-point differences averaged equally across models. Absolute country-by-language exact-match scores are reported in Table 48.

Country	Arabic Δ	Indonesian Δ	Persian Δ	Romanian Δ	Vietnamese Δ
Australia	-2.2	-0.2	-3.1	-0.1	-0.1
Brazil	-2.6	+0.7	-3.7	+2.3	+1.8
France	-0.4	+0.9	-1.7	+0.9	+0.3
Germany	-0.7	+2.1	-3.6	+1.7	+0.3
Netherlands	-0.7	+0.7	-4.3	+1.7	+1.1
Singapore	-2.4	-1.8	-4.7	-1.9	-2.0
South Korea	-2.3	-3.6	-4.3	-2.3	-2.6
Sweden	-1.2	-2.7	-3.2	-2.6	-0.4
United Kingdom	-5.9	-3.1	-6.6	-1.9	-0.6
United States	-5.7	-3.8	-7.0	-3.7	-2.1

F Additional Evidence for RQ 4: Auditable Language-Safety Assessment

F.1 Evidence Errors by Category and Model

Safety evidence reveals distinct precision–recall trade-offs. Sol has the highest category-agnostic recall (96.3%) but lower precision (70.9%) than Gemini (83.0%); DeepSeek reverses the pattern with 77.8% precision and 45.6% recall. Gemini is more balanced at 83.0% precision and 90.9% recall, yielding the highest agnostic F1 of 86.8%. These profiles distinguish higher false-positive tendencies from failures to recover relevant evidence.

Category assignment introduces an additional source of error. Gemini's category-agnostic F1 of 86.8% falls to 77.8% under strict matching, while Sol declines from 81.7% to 74.5%. The gap captures cases where a relevant subtitle line is recovered but assigned to the wrong lexical category. Count error provides an additional diagnostic: Gemini has the lowest count MAE among the evaluated models at 5.01 per film and lexical category on average.

Mild obscenity is difficult in both directions. Across models, mild obscenity has 26.3% recall, a 73.7% miss rate, and an 81.8% false-positive rate, yielding 20.3% category-level F1. Strong profanity is more lexically explicit, with 70.0% precision, 62.6% recall, and 62.8% F1. The model-by-category heat map shows how strongly individual systems depart from these averages.

Table 50: Evidence-grounded language-safety performance by model. Category-agnostic metrics measure recovery of safety-relevant subtitle indices regardless of lexical category, while strict F1 additionally requires the correct category assignment. Count MAE compares predicted and gold category counts and is macro-averaged across the four lexical categories. Best values are bold, and second-best distinct values are underlined.

Model	Agnostic P (%) ↑	Agnostic R (%) ↑	Agnostic F1 (%) ↑	Strict F1 (%) ↑	Count MAE ↓
GPT-5.6 Sol	70.9	**96.3**	81.7	74.5	7.16
Gemini 3.8 Flash	**83.0**	90.9	**86.8**	**77.8**	**5.01**
Claude Haiku 4.5	61.6	55.0	58.1	37.2	8.95
GPT-5 Nano	16.0	33.5	21.7	9.6	11.55
Gemini 3.5 Flash Lite	68.0	62.8	65.3	42.8	9.42
DeepSeek V4 Pro 1.6T	77.8	45.6	57.5	48.3	9.18
Qwen3 235B	68.8	33.4	45.0	29.6	10.27
Mistral 4 119B	59.7	42.4	49.6	31.7	7.78
Llama 4 Scout 17B	25.4	32.1	28.4	12.7	11.11

Table 51: Language-safety error profiles by lexical category, averaged across the nine reported English model runs. Arrows indicate the preferred direction: ↑ is higher-is-better and ↓ is lower-is-better. Miss rate measures gold safety-relevant evidence that was not recovered, false-positive rate measures unsupported predicted evidence, and cross-category FP measures evidence assigned to an incorrect lexical category. Teal shading is normalized within each metric column across categories. Best values are bold, and second-best distinct values are underlined at the displayed precision; ties share the same emphasis.

Category	Precision (%) ↑	Recall (%) ↑	F1 (%) ↑	Miss rate (%) ↓	False-positive rate (%) ↓	Cross-category FP (%) ↓	Count MAE ↓
Mild Obscenity	18.2	26.3	20.3	73.7	81.8	32.6	8.43
Crude Bodily Language	54.5	35.0	40.5	65.0	45.5	22.1	11.85
Religious profanity/exclamation	39.8	40.2	38.3	59.8	60.2	**14.8**	**7.56**
Strong Profanity	**70.0**	**62.6**	**62.8**	**37.4**	**30.0**	34.0	7.91

Table 52: English language-safety evidence performance by model and lexical category. Strict F1 requires both the correct subtitle index and lexical category, while recall measures recovery of gold evidence indices within each category. Strong, crude, mild, and religious abbreviate the four language-safety categories defined in the main text. Teal shading compares models within each category and metric. Best values are bold, and second-best distinct values are underlined at the displayed precision.

	Strict evidence F1 (%)				Evidence recall (%)			
Model	Strong	Crude	Mild	Religious	Strong	Crude	Mild	Religious
GPT-5.6 Sol	**99.6**	85.6	39.7	73.3	**100.0**	**86.6**	**72.1**	**94.4**
Gemini 3.8 Flash	99.1	**86.0**	**52.0**	**74.2**	99.2	82.3	59.1	87.8
Claude Haiku 4.5	68.5	33.5	17.7	29.1	72.2	23.2	24.9	22.4
GPT-5 Nano	22.4	9.4	2.0	4.8	34.9	12.3	8.4	4.1
Gemini 3.5 Flash Lite	61.2	40.2	22.0	47.9	45.9	34.1	28.5	56.7
DeepSeek V4 Pro 1.6T	68.2	48.4	22.6	54.1	52.3	33.3	20.5	48.4
Qwen3 235B	58.1	22.9	14.2	23.3	45.5	14.9	12.3	17.9
Mistral 4 119B	62.4	26.5	8.9	28.9	62.1	19.6	8.1	23.2
Llama 4 Scout 17B	25.7	12.5	3.5	9.2	51.6	8.9	2.6	7.2

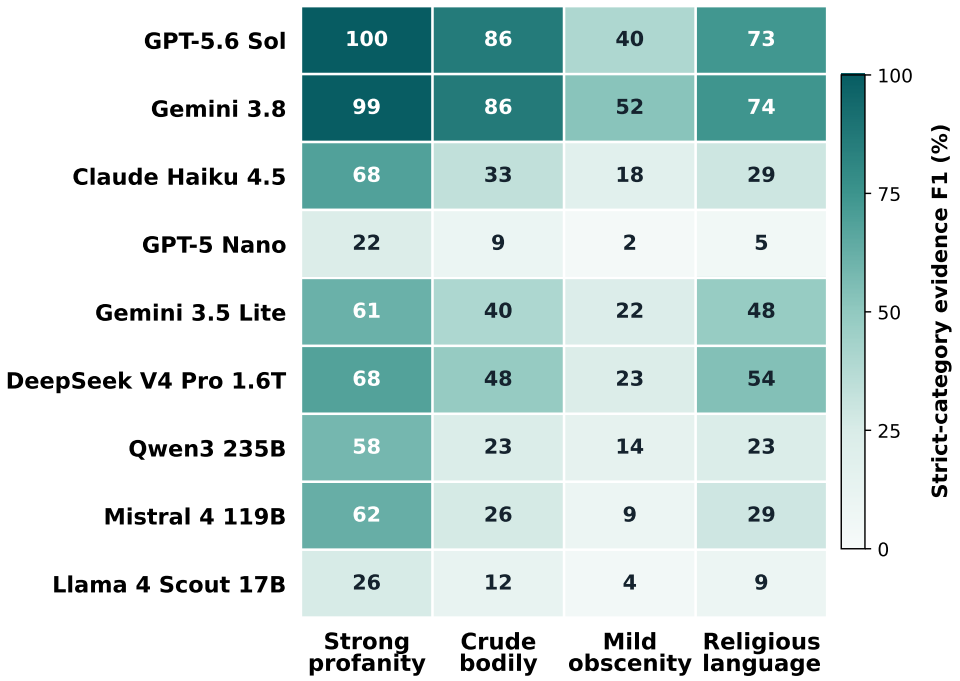


Figure 13: Strict-category evidence F1 by model and lexical category. Darker teal indicates higher F1. Strong profanity is easiest overall, while mild obscenity combines low recall with frequent false positives.

G Additional Analysis: Within-Family Scaling

Within-family scaling is task dependent. Holding model family and English input constant, Ministral narrative overall rises from 2.25 to 2.55 to 2.78 as model size increases from 3B to 8B to 14B, while age MAE falls monotonically from 2.42 to 1.92 to 1.41 years. Classification results are not monotonic: genre micro-F1 increases from 56.7% to 66.7% and then falls to 64.8%, while country-rating exact match moves from 22.5% to 36.6% to 35.6%. Within this family, additional scale therefore benefits narrative reconstruction and age prediction more consistently than genre or national-rating classification.

Table 53: English-only within-family scaling results for Ministral 3B, 8B, and 14B. Narrative performance improves and age MAE decreases monotonically with model size, while genre micro-F1 and country-rating exact match peak at 8B. Best values are bold, and second-best distinct values are underlined.

Model	Narrative ↑	Genre F1 (%) ↑	Age MAE ↓	Country EM (%) ↑
Ministral 3B	2.25	56.7	2.42	22.5
Ministral 8B	<u>2.55</u>	**66.7**	<u>1.92</u>	**36.6**
Ministral 14B	**2.78**	<u>64.8</u>	**1.41**	<u>35.6</u>

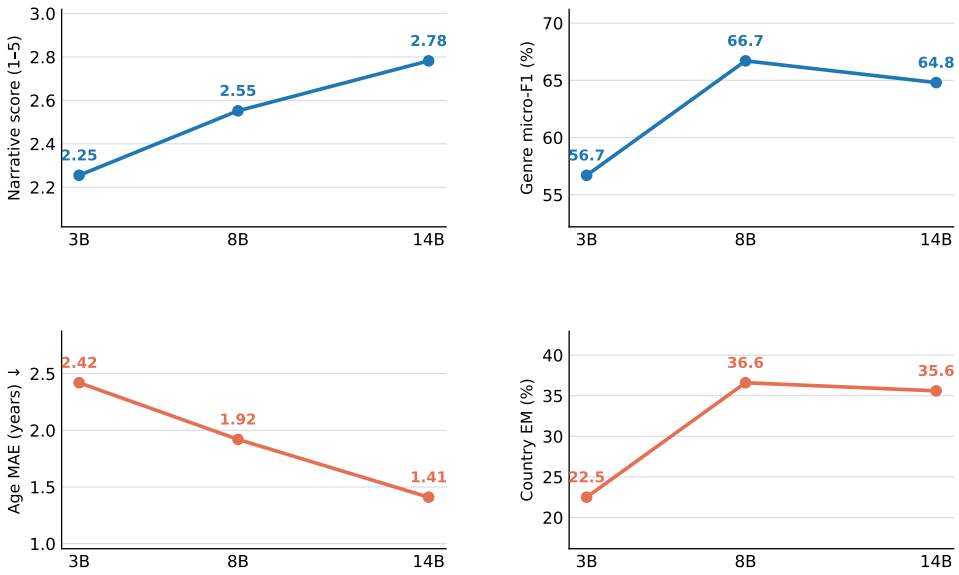


Figure 14: English-only Ministral scaling in original metric units. The panels report narrative overall, genre micro-F1, age MAE (lower is better), and national-rating exact match.

H Qualitative Case Studies

These cases illustrate distinct observable failure mechanisms and one within-film language effect; they are not estimates of population frequency. Text is drawn from saved outputs, and annotations compare it with benchmark gold fields or subtitle-index evidence. Film titles link to IMDb.

About Last Night (2014) 1h 40m English subtitles

Plot—gold. "Follow two couples as they journey from the bar to the bedroom and are eventually put to the test in the real world."

GPT-5.6 Sol—plot excerpt. "After cautious strangers Danny and Debbie meet through their uninhibited friends, a spontaneous hookup grows into love, cohabitation, and a shared dog."
Llama 4 Scout—plot. "The film follows Danny, a recently broken-up man, and his friend Bernie, a carefree and promiscuous womanizer, as they navigate their lives, relationships, and personal growth in **Chicago**."

Synopsis—gold evidence anchors. "Keller is ready to let Danny go, but Danny makes the first move and quits." Later, "Bernie and Joan have rekindled their relationship, even though it is mostly sexual."
Gemini 3.8 Flash—synopsis excerpts. "Danny despises his corporate restaurant supply job, eventually **getting fired after defying company policy** to supply a struggling neighborhood bar owned by **his late father's friend**, Casey." Later, "Bernie and Joan reconcile by accepting their unorthodox, fiercely combative dynamic and **move in together**."

Genre labels. Gold: Comedy, Romance. Gemini 3.8 Flash: Comedy, Romance, **Drama**.

Figure 15: A film-level case showing why fluent output is not necessarily faithful. Red marks unsupported location, event, relationship, or genre content; bronze marks an altered relationship outcome. The synopsis lines are verbatim excerpts, not complete synopses; the omitted passages are not scored here. All quotations and labels come from saved English-input records.

The Abyss (1989) 2h 20m English subtitles

Age suitability and national ratings. The same subtitle input produces different national label predictions. Teal cells match the gold label; pale red cells do not.

Target	Gold	GPT-5.6 Sol	GPT-5 Nano
Age suitability	**13+**	15+	17+
Australia	M	M	MA15+
Brazil	Livre	14	16
France	Tous publics	12	16
Germany	12	12	16
Singapore	PG13	PG13	M18
United Kingdom	15	15	15
United States	PG-13	PG-13	R

GPT-5 Nano—age rationale excerpt. "The content is mature and non-graphic, supporting an **adult-only level**." The gold age label is 13+, illustrating a conservative shift in the model's predicted threshold.

Figure 16: An English-input cultural case with seven of the ten country labels shown for legibility. The gold and predicted labels are the exact saved outputs, not translations into a common age scale; red denotes a mismatch within that country's own label space. Flags are rendered at one consistent display size. This case was chosen to illustrate a broad upward shift in age and national classifications, not to estimate its prevalence.

Alpha (2018) 1h 36m English subtitles

Key message—gold. "Empathy for other creatures leads to confidence and courage. Prehistoric humans made alliances with wolves, training them to help hunt for food."
GPT-5.6 Sol. "True strength grows through compassion, loyalty, and the courage to protect others."
Llama 4 Scout. "**Life is for the strong, and it is earned, not given.**" This shifts the film's emphasis from interspecies empathy to strength as a condition of survival.

Genre labels (multi-label output):

Gold	Adventure	Drama	
GPT-5.6 Sol	Adventure	Drama	**Family**

The highlighted label is outside the saved gold set.

Alien: Romulus (2024) 1h 59m English subtitles

Subtitle-grounded language-safety assessment. Two direct strong-profanity cues in the same subtitle track are selected by GPT-5.6 Sol but omitted by Qwen3 235B, illustrating why an auditable warning must expose its supporting dialogue:

Line	English subtitle excerpt	Gold / GPT-5.6 Sol	Qwen3 235B
#214	"Bjorn, shut the fuck up."	Selected / selected	**Not selected**
#332	"for fuck's sake."	Selected / selected	**Not selected**

Both omissions are visible directly in the saved strong-profanity evidence-index lists; the quoted lines are excerpts from the corresponding subtitle blocks.

Figure 17: Two complementary English-input cases. Violet marks a thematic reversal, red a predicted genre outside the saved gold set, and bronze a missed explicit evidence line. Gold labels, subtitle text, and model outputs are taken from saved sample-level records. The examples illustrate mechanisms; they are not estimates of error frequency.

***The Master* (2012) 2h 18m GPT-5.6 Sol**

One film, one model, two subtitle languages. Gold labels stay fixed while the input changes from English (en) to Vietnamese (vi). Teal marks exact matches; pale red marks mismatches. In this case, Vietnamese input changes four country predictions from mismatches to matches and corrects the age prediction.

Target	Gold	English (en)	Vietnamese (vi)
Age suitability	17+	18+	17+
Australia	MA15+	R18+	MA15+
Brazil	14	18	16
France	Tous publics	16	16
Germany	12	16	16
Netherlands	12	16	16
Singapore	M18	R21	M18
South Korea	19	19	19
Sweden	15	15	15
United Kingdom	15	18	15
United States	R	NC-17	R

Country exact matches: English 2/10; Vietnamese 6/10. The four persistent mismatches are retained in the table to show the full ten-country outcome, not only the changed labels.

Figure 18: An illustrative within-film cross-lingual contrast from saved GPT-5.6 Sol outputs. Only the subtitle-input language changes; the gold age and national-rating labels are identical in both runs. This selected case makes a possible cultural benefit of multilingual input tangible, while the aggregate language comparison in the main paper supports the population-level analysis.

***Carol* (2015) 1h 58m English subtitles**

Plot—gold. "An aspiring photographer develops an intimate relationship with an older woman in 1950s New York."

Gemini 3.8 Flash—plot excerpt. "In 1950s New York, an aspiring young photographer working at a department store meets an elegant, sophisticated older woman navigating a painful divorce. A magnetic connection develops between them" [remainder omitted].

Llama 4 Scout—complete plot. "A young woman, Therese, meets **an older married man, Gray, in a taxi**. They strike up a conversation, and Therese is drawn to Gray's world. As they navigate their complicated relationships and desires, Therese becomes embroiled in **Gray's marriage to Carol**, a woman with a troubled past."

What changed? The reference synopsis begins with Therese meeting *Carol* in a department store; Carol is the older woman and Therese's romantic counterpart. The highlighted taxi meeting and marriage invert that central relationship. The narrative evaluator assigns Llama's plot **1/5** for this sample.

Figure 19: A visually fluent but wrong-premise plot, contrasted with a model that preserves the central relationship. Red spans are exact phrases from Llama's saved output; Gemini's bracketed omission is editorial and not part of its output. The counterexample demonstrates a failure mode, not its prevalence.

I LIMITATIONS, INTENDED USE, AND DATA STATEMENT

I.1 LIMITATIONS

Subtitle variation and cross-lingual interpretation. Subtitle tracks vary across releases and translations. We address this through structural, temporal, density, language-consistency, and manual checks, and evaluate all six languages over the same matched films. The resulting cross-lingual comparisons capture model behavior under each subtitle-language condition, encompassing translation, tokenization, subtitle density, and model-language proficiency.

Coverage and benchmark scope. CineSubBench prioritizes complete matched coverage across six subtitle languages and ten national classification systems, yielding 1,012 films for controlled MultiX comparison. This design supports direct comparison across tasks, languages, and cultural settings rather than population-level characterization of global cinema.

Reference labels and normalization. Age recommendations and national ratings retain the conventions of their respective sources and institutions. National systems are preserved as separate ordered label spaces, with historical or variant labels normalized only where needed for consistent evaluation. Narrative references provide a common semantic anchor while permitting valid variation in wording and emphasis.

Language-safety scope. The language-safety task provides auditable lexical-evidence assessment over four categories in the English subtitle track, enabling line-level evaluation of both evidence recovery and category assignment. Extending the evidence taxonomy and multilingual annotations offers a natural direction for future benchmark development.

Source exposure and prior film knowledge. Public film information can appear in model pretraining data. We isolate this factor by withholding titles, release years, cast, and reference answers from standard CineSubBench inputs. Appendix B.6 separately measures the signal available from title/year cues and sensitivity to character-name perturbations, distinguishing prior film information from subtitle-conditioned generation.

Modality scope. CineSubBench intentionally uses subtitles as a common film-length linguistic interface that supports consistent evaluation across API-based and open-weight LLMs. It measures narrative and cultural understanding from temporally ordered subtitle evidence, while visual action, cinematography, music, gestures, and other non-speech signals form a complementary multimodal evaluation setting.

I.2 INTENDED USE

CineSubBench is intended for research on long-form film understanding, multilingual subtitle processing, culturally situated audience assessment, and auditable language-safety evaluation. Its timestamped multilingual tracks can also support research on subtitle alignment, cross-lingual retrieval, temporal reasoning, summarization, and accessibility-oriented subtitle systems. Country-specific ratings should remain associated with their originating national systems, while language-safety evidence is intended for transparent, human-reviewable assessment rather than automated censorship.

I.3 DATA STATEMENT

CineSubBench contains structured records for 1,012 films, including source metadata, normalized task labels, six timestamped subtitle tracks per film, subtitle-indexed language-safety evidence, and benchmark task definitions. Source links and provenance fields are retained for traceability, with construction, verification, sanitization, and field provenance documented in Appendix A. Release and redistribution follow the requirements of the respective source providers.