Beyond Factual QA: Mentorship-Oriented Question Answering over Long-Form Multilingual Content

Parth Bhalerao **Ruiwen Guan** **Diola Dsouza** **Oana Ignat**
Department of Computer Science and Engineering, Santa Clara University
Santa Clara, CA, USA
{pbhalerao,oignat}@scu.edu

Abstract

Question answering systems are usually evaluated for factual correctness, but many real world settings such as education and career guidance require mentorship responses that offer reflection, context, and guidance. We introduce MentorQA, the first multilingual dataset and evaluation framework for mentorship focused question answering from long form videos, with nearly 9,000 QA pairs from 180 hours of content across four languages. We define evaluation dimensions that measure fluency, clarity, alignment, and mentorship value, and use them to compare Single Agent, Dual Agent, RAG, and Multi Agent QA pipelines under controlled conditions. Multi Agent pipelines produce the strongest mentorship responses, with larger gains for complex topics and lower resource languages. We also study LLM based evaluation and find substantial variation in alignment with human judgments. Overall, MentorQA establishes mentorship focused QA as a distinct research problem and provides a benchmark for studying agentic architectures and evaluation design in educational AI. The dataset and evaluation framework are released at https://aim-scu.github.io/MentorQA.

1 Introduction

Question answering (QA) systems are typically evaluated on factual correctness (Mulla and Gharpure, 2023; Fu et al., 2024). However, many real world settings such as education, career guidance, and personal development require *mentorship*: answers that provide context, reflection, and guidance rather than isolated facts (Ma and Ma, 2019; Kovács, 2016). Existing QA benchmarks rarely capture this need, especially for multilingual and long form content where useful guidance is distributed across extended narratives.

This gap limits the study of mentorship focused QA in two ways. First, prior QA research has largely emphasized short form inputs, factual accuracy, and monolingual evaluation (Faraby et al., 2024). Multilingual QA benchmarks assess fluency and correctness, but seldom measure whether answers provide learning value or actionable guid-

English | Career/Education | Multi-Agent **Q:** What should a student consider when choosing a research topic that aligns with their interests? **A:** When selecting a research topic that aligns with their interests, a student should consider several key factors: 1. **Interest and Passion**: The student must genuinely... **Scores:** QFlu 5, AFlu 5, QClar 4, AClar 5, QA-Algn 5, QMen 5, AMen 5
Chinese | Education | Single-Agent **Q:** 千禧世代的出生年份是什麼? *(What are the birth years of the Millennial generation?)* **A:** 千禧世代的出生年份是 1980 年至 1994 年或 1995 年。 *(The birth years of the Millennial generation are 1980 to 1994 or 1995.)* **Scores:** QFlu 3, AFlu 4, QClar 5, AClar 2, QA-Algn 4, QMen 1, AMen 1
Hindi | Business | Dual-Agent **Q:** भारत की ऐतिहासिक रूप से धनवान था तो उसके व्यापार का क्या रूप था? *(If India was historically wealthy... how was the nature of trade?)* **A:** भारत ने चानक्य के समय से बीटा को बोर्ड ट्रेड के एक अच्छे देश के रूप में जाना था।..... डाट बिक था जो तब विश्व की दुनिया का नक्षा था। *(From Chanakya's time, India was known as.. and carried out trade with..)* **Scores:** QFlu 3, AFlu 3, QClar 2, AClar 3, QA-Algn 5, QMen 1, AMen 3
Romanian | Mental Health | RAG **Q:** Ce implică muncă prea puțin stimulantă și ce efecte are asupra sistemului nervos? *(What does work that is too little stimulating imply..... and affects nervous system?)* **A:** Muncă prea puțin stimulantă implică simțirea de plictisire și blazare, ceea ce poate duce la o solicitare excesivă... *(Work that is not stimulating enough involves feelings of boredom.... lead to excessive strain on the nervous system, causing.....)* **Scores:** QFlu 2, AFlu 4, QClar 2, AClar 4, QA-Algn 1, QMen 4, AMen 4

Table 1: MentorQA multilingual examples across languages, topics, and generation models. Scores denote Fluency (Q/AFlu), Clarity (Q/AClar), QA Alignment (QA-Algn), and Mentorship (Q/AMen) on a 1 to 5 Likert scale.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 20086–20107
October 24-29, 2026 ©2026 Association for Computational Linguistics

ance. Second, while recent language models have advanced open domain QA (Guo et al., 2024), it remains unclear how architectural choices such as retrieval augmentation and agentic coordination affect mentorship quality across languages with different discourse norms and resource availability.

Contributions. First, we introduce MentorQA, the first multilingual dataset and evaluation framework for mentorship focused question answering from long form videos. The dataset contains nearly 9,000 QA pairs from 180 hours of mentorship content across four languages. **Second, we define evaluation dimensions that capture clarity, alignment, and mentorship value, and compare Single Agent, Dual Agent, RAG, and Multi Agent QA architectures under controlled settings.** Our results show that Multi Agent pipelines consistently achieve higher mentorship quality, with particularly strong gains for complex topics and lower resource languages. **Finally, we examine the reliability of automated LLM based evaluation across languages and metrics**, finding substantial variation in alignment with human judgments. Throughout this work, mentorship focused QA refers to the task objective, while mentorship value refers to the learning value and actionable guidance captured by our Question Mentorship and Answer Mentorship metrics. **Together, our findings establish mentorship focused QA as a distinct research problem and provide a multilingual benchmark for studying agentic architectures and evaluation methods in educational AI.**

2 Related Work

Educational and Mentorship QA from Long Transcripts. Recent work extracts questions and answers from long educational recordings through lecture summaries (Chen and Yen, 2024), video based QA context selection (Yu et al., 2026), and segment level retrieval for podcasts and meeting transcripts (Elaryan, 2022; Prasad et al., 2023). Other work studies personalized QA for language learning assistants (Sammoudi et al., 2025) and multi agent debate for online education (Du et al., 2025). These systems support educational content, but they mainly optimize factual correctness, content coverage, or single modeling settings. In contrast, mentorship focused QA emphasizes guidance, reflection, and learning value in long form multilingual content, which remains underexplored.

Agentic and RAG Based QA Pipelines. Agentic models support complex workflows for long document understanding and QA by coordinating specialized agents for reading, reasoning, and generation (Wang et al., 2025; Saadaoui and Alonso, 2025), with recent surveys summarizing common collaboration mechanisms (Tran et al., 2025). Retrieval augmented generation has also been used for domain specific QA, including health risk assessment (Meng et al., 2025) and podcast or meeting analysis (Aquilina et al., 2023; Zhu et al., 2025). However, these approaches are rarely compared under controlled conditions or evaluated for mentorship quality. We compare Single Agent, Dual Agent, RAG, and Multi Agent pipelines with a fixed base model to isolate the effect of architecture on mentorship focused QA.

Multilingual QA and Metric Based Evaluation. Prior work has studied question generation evaluation and multilingual QA through automatic QG metrics (Nema and Khapra, 2018), reference free and answer based metrics such as RQUGE (Mohammadshahi et al., 2023) and QG Eval (Fu et al., 2024), LLM as judge evaluation (Wolfe, 2024), and multilingual QA benchmarks (Ushio et al., 2023; Moreno-Cediel et al., 2024; Ruder and Sil, 2021; Asai et al., 2021). Yet automatic metrics often correlate weakly with human judgments for task oriented dimensions such as usefulness or guidance (Fu et al., 2024). We therefore introduce mentorship focused metrics that capture learning value and support comparison between human and automatic evaluation.

3 The MentorQA Dataset

We introduce MentorQA, to our knowledge the first QA dataset focused on mentorship. Unlike factual QA benchmarks, MentorQA centers on guidance, reflection, and practical insight. We define mentorship value as information that helps learners grow through advice, perspective, or actionable guidance. The dataset contains 8,990 QA pairs from 120 long form mentorship videos across four languages and six topics.

Video Collection and Pre Processing. We collect podcast style conversations and panel discussions from YouTube that provide reflective, experience driven guidance. Each video averages over 1.5 hours, resulting in 180 hours of multilingual men-

torship content. We process only the audio track and transcribe it with Whisper, which detects the source language and produces transcripts for QA generation.

Languages. The dataset covers *English*, *Chinese*, *Hindi*, and *Romanian*, selected for typological and cultural diversity while supporting reliable annotation. Videos and QA pairs are balanced across languages and evaluated by native or expert speakers.

Topic Selection across Languages. Videos cover six mentorship topics[1]: *Entrepreneurship, Education, Finance, Mental Health, Personal Growth*, and *Physical Health*. Topic labels are assigned using *Qwen2.5 7B* (Yang et al., 2025) and verified by expert annotators. English videos focus on academic and professional mentorship, including research directions, publishing, and PhD pathways through ACL Mentorship. Romanian videos cover education, personal development, and mental health through conversations with authors, coaches, and athletes from Mind Architect. Chinese videos focus on experience driven mentorship in mental health, career growth, finance, and physical well being from Drinking Library. Hindi videos focus on career advancement, business, and mental or physical health from The Ranveer Show.

QA Generation. We generate QA pairs from each transcript using *Single Agent*, *Dual Agent*, *RAG*, and *Multi Agent*. Each model produces about 20 mentorship focused QA pairs per video, giving about 80 QA pairs per video and 8,990 pairs overall (Table 2). Multilingual examples are shown in Table 1, and model architectures are described in Section 4.

Quality Assurance. Each QA pair is evaluated along four dimensions: *Fluency*, *Clarity*, *QA Alignment*, and *Mentorship Value*. Human evaluation uses three annotators per language, complemented by nine LLM based judges for scalable analysis and comparison with human judgments. Evaluation details are in Section 5.

PII Anonymization. Although the source data are public, we reduce downstream propagation of personally identifiable information. We combine named entity recognition based on XLM RoBERTa, structured PII detection and anonymization via Presidio, LLM prompts for implicit identifiers, and human review for lower resource languages and culturally specific entities.

Scalability. MentorQA is released as an extensible open source resource. Although the current version covers four languages, selected channels, and one video platform, the codebase and documentation support expansion to additional languages and content sources.

[1]All the mentorship videos are in Section A.1.

Dataset Scale and Coverage	
Videos (long form)	120
Total duration (hours)	180
Languages	4
Mentorship topics	6
QA generation models	4
Total QA pairs	8,990
Evaluation and Quality Control	
Evaluation dimensions	4
LLM judges	9
Human annotators	12
Mentorship Topics (# QA pairs)	
Personal Growth	4,450
Career & Education	2,689
Mental Health	741
Business & Entrepreneurship	392
Physical Health	431
Finance	287

Table 2: MentorQA statistics.

4 Mentorship focused QA Generation

Mentorship focused QA from long form conversational videos requires more than long context reasoning and topic tracking. Mentorship value is unevenly distributed across transcripts: brief segments with reflection, advice, or experience driven insight can be more useful than longer descriptive passages. Effective systems must therefore identify valuable guidance and allocate questions by mentorship quality rather than transcript length.

We study four QA generation pipelines of increasing structure: *Single Agent*, *Dual Agent*, *Retrieval Augmented Generation (RAG)*, and *Multi Agent*. These pipelines introduce topic awareness, answer grounding, and explicit control over mentorship quality and question allocation. Across all pipelines, we use the same base model, *Qwen 2.5 7B Instruct 1M* (Yang et al., 2025), in long context mode, ensuring that differences reflect architecture rather than model capacity. Full prompts are provided in subsection B.1.

The four pipelines are designed as distinct architectures to isolate architectural effects. Topic aware chunking, retrieval based grounding, and question scoring with filtering represent separate design dimensions; combining them would ob-

scure which component drives performance differences.

4.1 Single Agent

The Single Agent baseline processes the full transcript with one long context LLM. The model is prompted as an expert educational analyst to generate 20 mentorship focused QA pairs that cover the conversation, avoid trivial or overly local questions, and provide educational value. This baseline relies only on the long context capability of the LLM, without segmentation, retrieval, or coordination.

4.2 Dual Agent

The Dual Agent pipeline separates topic segmentation from QA generation to reduce over concentration in dense or early transcript regions. A *Chunking Agent* partitions the transcript into coherent topic based segments while preserving the original content and adapting segment length to topic duration. A *QA Agent* then generates a fixed number of mentorship focused QA pairs per segment using the same prompt as the Single Agent baseline. This improves coverage across the conversation and better aligns questions with mentorship themes.

4.3 Retrieval Augmented Generation (RAG)

RAG separates question generation from answer grounding. The model first generates a fixed set of diverse mentorship focused questions from the full transcript. Each question is then answered using relevant transcript segments retrieved through similarity search. We build a multilingual retrieval index with BGE M3 (Chen et al., 2024), which chunks long transcripts and supports all four dataset languages. Retrieved chunks are provided to a second LLM query to generate the final answer, enabling global question coverage while grounding answers in relevant transcript evidence.

4.4 Multi Agent

Dual Agent improves coverage through topic based segmentation, and RAG grounds answers through retrieval, but both use single pass question generation. *In practice, mentorship quality varies substantially across topics: uniform question allocation can over represent weak segments and under represent parts of the conversation rich in guidance.* We therefore introduce a Multi Agent framework that scores mentorship quality and allocates questions across topics based on estimated value. As shown in Figure 1, the pipeline uses specialized agents for topic discovery, question exploration, quality assessment, allocation, and answer generation.

The process starts with the same *Chunking Agent* used in Dual Agent, which performs dynamic topic based segmentation and assigns each segment a concise title. For each segment, a *Question Brainstorming Agent* over generates diverse mentorship relevant candidate questions. A *Scorer Agent* assigns each candidate question q an integer mentorship score, $score(q) \in \{1, \ldots, 10\}$. We then compute segment strength as:

$$S(i) = \mu(q) - \sigma(q), \tag{1}$$

where $\mu(q)$ and $\sigma(q)$ denote the mean and standard deviation of question scores within segment i. This favors segments with consistently high mentorship value and penalizes noisy or uneven question quality. Given a target of $T = 20$ questions per video, matching the generation budget used by the other architectures, we distribute this budget across segments proportionally to their strength scores. Specifically, the quota for segment i is:

$$\text{quota}(i) = \text{round}\left(\frac{S(i)}{\sum_{j \in \mathcal{S}} S(j)} \times T\right), \tag{2}$$

ensuring that stronger segments, those with higher density of mentorship relevant content, contribute more questions to the final set. Within each segment, only the top ranked questions are retained. A *Justifier Agent* explains question selection or rejection to improve transparency and reduce redundancy. Finally, an *Answering Agent* generates grounded final answers from the corresponding segment with mentorship oriented explanations. Overall, this design enables broader question exploration, explicit mentorship quality control, and adaptive allocation to the most valuable parts of long form conversations.

4.5 Evaluation Metrics

We build on QG Eval (Fu et al., 2024), which separates linguistic and task oriented dimensions, and extend it to capture QA alignment and mentorship value. Fluency and Clarity are evaluated separately for questions and answers to support fine grained analysis.

Linguistic Metrics. We assess surface level language quality for both questions and answers:

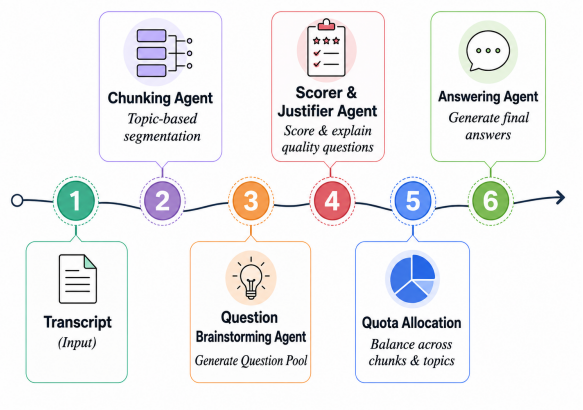


Figure 1: Multi Agent QA generation pipeline with topic segmentation, question exploration, quality scoring, quota based allocation, and grounded answer generation.

- **Fluency**: *The question or answer is grammatically correct and free of language errors.*
- **Clarity**: *The question or answer is easy to understand, specific, and unambiguous.*

Task Oriented Metrics. We assess whether QA pairs provide alignment and mentorship effectiveness:

- **QA Alignment**: *The question and answer pair demonstrates proper alignment, where the answer adequately satisfies what the question asks.*
- **Question Mentorship**: *The question elicits learning, guidance, advice, or insights that would benefit from mentor expertise.*
- **Answer Mentorship**: *The answer provides guidance, wisdom, practical advice, or insights that help the reader learn or grow.*

Together, these metrics capture linguistic quality, semantic alignment, and mentorship effectiveness.

Rating Scale. All metrics are rated on a 5 point Likert scale from 1, strongly disagree, to 5, strongly agree. Pilot studies showed that the 3 point scale used in prior work (Fu et al., 2024) was insufficient to capture nuanced quality differences and resulted in lower inter annotator agreement.

4.6 Human Evaluation

We recruit 12 annotators in four language groups, with three native or fluent speakers per language. Annotators rate a stratified sample of 240 QA pairs (15 per model $\times$ 4 models $\times$ 4 languages), with three independent ratings per pair. The sample is drawn from 20 videos verified by human reviewers to contain high mentorship value, with five videos selected per language. Although this is a subset of the full dataset, the stratified design ensures equal representation across model and language conditions, enabling reliable comparisons. Larger scale annotation was not feasible because mentorship oriented QA requires native speaker judgment and domain familiarity. An example annotation interface is provided in Section C.

Inter Annotator Agreement. We measure agreement using Gwet's AC2, computed separately for each metric across languages and models. We use Gwet's AC2 rather than Krippendorff's alpha because alpha can give misleading estimates under skewed rating distributions common in AI system evaluation (Battisti and Ebling, 2024; Pradhan et al., 2025).

4.7 Automatic Evaluation

Recent work shows that LLM based evaluators outperform traditional Natural Language Generation (NLG) metrics and provide a scalable alternative to human evaluation (Fu et al., 2024; Kocmi and Federmann, 2023). To study automatic evaluation for mentorship focused QA, we evaluate all QA pairs using nine multilingual LLM judges.

We select nine open source or publicly documented multilingual judges for reproducibility: Qwen2.5 7B (Yang et al., 2025), Command R 7B (Cohere, 2024), Mistral 8B Instruct (Mistral AI, 2024), Llama3.2 11B (Meta, 2024), Qwen3 8B (Team, 2025), Gemma3 12B IT (Google DeepMind, 2025), Pangea 7B (Yue et al., 2025), AyaExpanse 8B (Dang et al., 2024), and Phi 4 14B (Abdin et al., 2024). Only two of the nine judges belong to the Qwen family, while the remaining judges span several independent model families.

LLM Human Agreement. We compute Gwet's AC2 between LLM scores and human annotations for each metric, language, and QA generation model to quantify when automatic evaluation aligns with human judgments.

5 Evaluation Results

We evaluate mentorship focused QA generation along four axes: reliability of mentorship focused evaluation, scalability through LLM judges, effectiveness of QA generation models, and consistency across languages and topics.

Reliability of Mentorship Focused Evaluation. Human annotations show the highest agreement

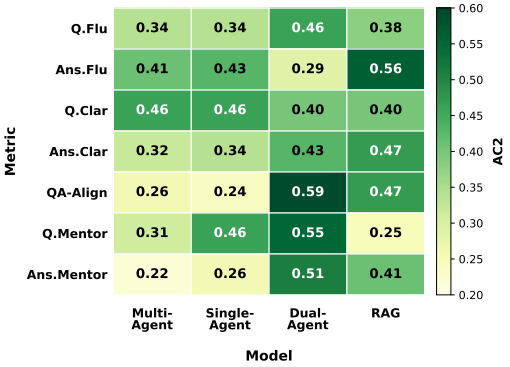


Figure 2: IAA Scores.(Averaged across Languages)

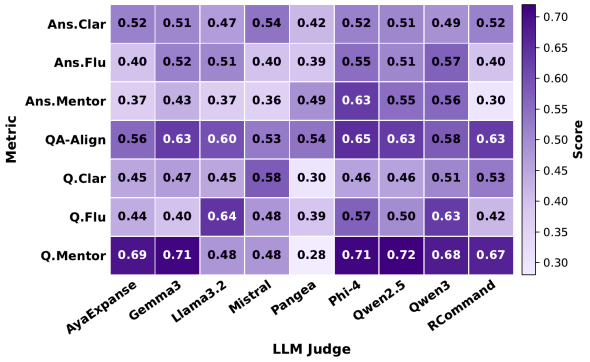


Figure 3: LLM-human agreement scores, with Qwen models, Phi-4, and Gemma3 aligning best.

for linguistic metrics: *Fluency* (Q: 0.38, A: 0.42) and *Clarity* (Q: 0.43, A: 0.39). Task oriented metrics also show stable moderate agreement: *QA Alignment* (0.39) and *Mentorship* (Q: 0.39, A: 0.35), reflecting the interpretive nature of mentorship quality. Agreement varies by model and language (Figure 2, language-averaged). *Dual Agent shows the highest agreement* (0.46), followed by RAG (0.42), Single Agent (0.36), and Multi Agent (0.33). Across languages, agreement is highest for Romanian (0.67) and lower for Chinese (0.37), English (0.28), and Hindi (0.26) (more in Section C). These patterns show that mentorship focused evaluation is sensitive to model behavior and linguistic context, motivating culturally aware and mentorship specific evaluation.

Human LLM Agreement in Automatic Evaluation. To assess scalability, we compare LLM based judges with human annotations in Figure 3. Agreement is highest for task oriented metrics, with Qwen2.5, Qwen3, Phi 4 and Gemma3 reaching substantial agreement with human ratings (0.55 to 0.72), while linguistic metrics show moderate but stable alignment (0.40 to 0.57). *Qwen family models and Phi 4 achieve the strongest agreement across metrics and languages*, including lower resource settings (results per language in Section C). However, agreement varies substantially across judges, motivating language and metric aware evaluation with multiple LLM judges rather than a single evaluator.

Transcript Quality. Since all pipelines use automatic transcripts, we assess transcript quality as an upstream factor. Whisper Large v3 (OpenAI, 2023) reports error rates of 9.3 for English, 10.8 for Romanian, 12.8 for Chinese, and 16.9 for Hindi. For Chinese, the reported value is computed at the character level, while the remaining languages use word level error rates. All four architectures receive identical transcripts, so transcription errors do not confound cross architecture comparisons. Annotators for English, Romanian, and Chinese did not flag transcription induced QA errors. Hindi annotators noted occasional issues consistent with its higher WER, but replacing Qwen2.5 7B with Gemma3 27B for Hindi (Section 5) substantially improves agreement, suggesting base model capacity rather than transcript quality is the main bottleneck.

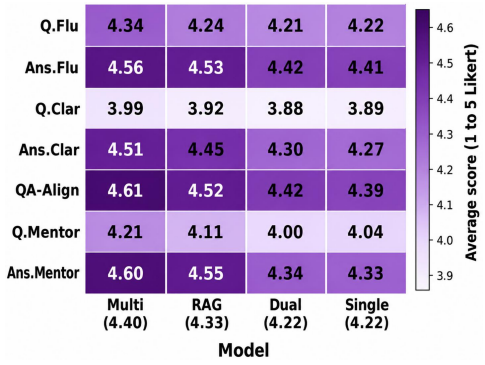


Figure 4: QA generation performance across models and metrics (1 to 5 Likert). Models are sorted by mean.

Comparison of QA Generation Architectures. The Multi Agent model performs best overall, ranking highest across all evaluation dimensions with a mean score of 4.40, as shown in Figure 4. RAG follows closely at 4.33 and improves over Single Agent and Dual Agent, especially on *QA Alignment*, *Question Mentorship*, and *Answer Mentorship*. Single Agent and Dual Agent perform nearly identically at 4.22, suggesting that topic aware chunking alone yields limited gains. Differences are largest for *Mentorship* and *QA Alignment*, while *Fluency* remains similar across models. Detailed cross model and LLM judge results are in Section C.2.

Statistical Significance Analysis. We test architectural differences in human ratings using pro-

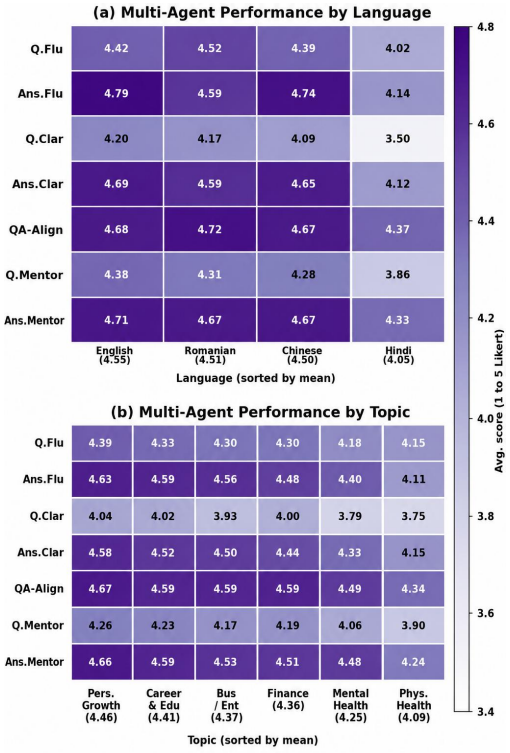


Figure 5: Multi Agent performance across (a) language and (b) topic. Topics and languages are sorted by mean.

Comparison	Q. Men.	A. Men.
Multi vs. Single	1.88 [1.15, 3.08]*	8.88 [5.66, 13.94]**
Multi vs. Dual	4.38 [2.58, 7.44]**	10.80 [6.38, 18.28]**
Multi vs. RAG	1.57 [0.99, 2.47]	1.74 [1.12, 2.71]*

Table 3: Pooled proportional odds OLR results across all languages. Cells report odds ratios with 95% confidence intervals. Q. Men. denotes Question Mentorship and A. Men. denotes Answer Mentorship. Significance markers use Holm adjusted p values across the six planned contrasts: $^{*}p < 0.05$, $^{**}p < 0.001$.

portional odds Ordinal Logistic Regression (OLR) over the full 1 to 5 rating scale. This analysis preserves the ordering of all response categories rather than reducing the ratings to a single threshold. For the pooled analysis, the model includes language, annotator, and source video fixed effects. The language specific models include annotator and source video fixed effects. Standard errors are clustered by QA item because each QA pair receives three independent human ratings. We report odds ratios (ORs), where an OR greater than 1 indicates higher odds of receiving a better rating for Multi Agent relative to the comparison architecture. Holm correction is applied across the six planned Multi Agent contrasts for the two mentorship outcomes within each analysis. Table 3 reports the pooled results across all languages.

Across all languages, the global Wald tests show significant architectural differences for both *Question Mentorship* ($\chi^2(3) = 30.33$, $p < 0.001$) and *Answer Mentorship* ($\chi^2(3) = 123.26$, $p < 0.001$).

For *Question Mentorship*, Multi Agent receives significantly higher ratings than Single Agent (OR$=$ 1.88, 95% CI $[1.15, 3.08]$, Holm adjusted $p = 0.037$) and Dual Agent (OR$=$ 4.38, 95% CI $[2.58, 7.44]$, Holm adjusted $p < 0.001$). The difference between Multi Agent and RAG is smaller and does not remain significant after correction (OR$=$ 1.57, 95% CI $[0.99, 2.47]$, Holm adjusted $p = 0.055$).

For *Answer Mentorship*, Multi Agent has substantially higher odds of receiving better ratings than Single Agent (OR$=$ 8.88, 95% CI $[5.66, 13.94]$, Holm adjusted $p < 0.001$) and Dual Agent (OR$=$ 10.80, 95% CI $[6.38, 18.28]$, Holm adjusted $p < 0.001$). Multi Agent also significantly outperforms RAG in the pooled analysis (OR$=$ 1.74, 95% CI $[1.12, 2.71]$, Holm adjusted $p = 0.037$). These results provide strong ordinal evidence that the Multi Agent architecture improves the mentorship quality of generated answers.

Because the pooled effect is particularly strong for *Answer Mentorship*, we further examine whether this pattern is observed across individual languages. The global architecture effect is significant for Answer Mentorship in English ($\chi^2(3) = 40.15$, $p < 0.001$), Chinese ($\chi^2(3) = 22.25$, $p < 0.001$), Hindi ($\chi^2(3) = 26.00$, $p < 0.001$), and Romanian ($\chi^2(3) = 40.98$, $p < 0.001$). As shown in Table 4(a) and Table 4(b), Multi Agent receives significantly higher Answer Mentorship ratings than both Single Agent and Dual Agent in every language after Holm correction. The contrasts with RAG are smaller and do not reach significance after correction within the individual language analyses.

The language specific analysis therefore reinforces the pooled result for Answer Mentorship. Although the estimated effect size varies across languages, Multi Agent shows a consistent and statistically significant advantage over Single Agent and Dual Agent. Importantly, these results are based entirely on human ratings, providing independent support for the mentorship gains observed in the broader evaluation and showing that the effect cannot be explained solely by a preference for longer responses among LLM judges.

Comparison	OR	95% CI	Holm p
English			
Multi Agent vs. Single	7.68	[3.57, 16.50]	< 0.001
Multi Agent vs. Dual	45.46	[12.63, 163.62]	< 0.001
Multi Agent vs. RAG	2.17	[0.91, 5.14]	0.158
Chinese			
Multi Agent vs. Single	3.53	[1.55, 8.07]	0.011
Multi Agent vs. Dual	6.16	[2.44, 15.54]	< 0.001
Multi Agent vs. RAG	0.96	[0.36, 2.53]	0.932

Table 4: (a) Language specific proportional odds OLR results for Answer Mentorship in English and Chinese. Each comparison reports Multi Agent relative to the indicated architecture. Holm adjusted p values are computed across the six planned mentorship contrasts within each language.

Comparison	OR	95% CI	Holm p
Hindi			
Multi Agent vs. Single	7.19	[2.74, 18.89]	< 0.001
Multi Agent vs. Dual	6.14	[2.74, 13.79]	< 0.001
Multi Agent vs. RAG	2.59	[1.18, 5.71]	0.072
Romanian			
Multi Agent vs. Single	63.81	[17.79, 228.93]	< 0.001
Multi Agent vs. Dual	17.81	[3.75, 84.55]	0.001
Multi Agent vs. RAG	1.74	[0.65, 4.68]	0.280

Table 4: (b) Language specific proportional odds OLR results for Answer Mentorship in Hindi and Romanian. Each comparison reports Multi Agent relative to the indicated architecture. Holm adjusted p values are computed across the six planned mentorship contrasts within each language.

Qualitative Analysis. Tables 5, 6, 7, and 8 show representative QA pairs across the four languages. Single Agent responses are often brief and factual, while Multi Agent responses provide more structured, point-wise actionable guidance, which human annotators rated as more helpful across languages.

Language Evaluation. Using the best performing Multi Agent model with scores averaged over all LLM judges, we analyze language effects in Figure 5(a), with cross model comparisons in Section C. *English performs best overall (4.55)*, while *Romanian (4.51) and Chinese (4.50)* remain comparable on *QA Alignment* and *Answer Mentorship*. *Hindi scores lower overall (4.05)*, especially on *Question Clarity* (3.50) and *Question Mentorship* (3.86), consistent with Devanagari tokenization and training limitations, as well as MMLU results where Hindi trails English (large-traversaal, 2025). However, Multi Agent narrows this gap: compared

Model	Question / Answer	Q.Men	A.Men
Single	**Q:** Can generative models be used for creating content in low-resource languages? **A:** Generative models do not work well for low-resource languages, generating only garbage... should be used cautiously by native speakers only.	4	2
Dual	**Q:** What are some areas of research that the speaker admires? **A:** The speaker admires research in Guarani papers... and OCR for low-resource and parsing papers.	3	3
RAG	**Q:** Can you provide examples of successful low-resource language projects? **A:** Examples include text-to-speech systems... for Tunisian Arabic and Korean... surprisingly low-resource despite more digital resources.	4	3
Multi	**Q:** What strategies overcome challenges collaborating with native speakers of low-resource languages? **A:** (1) *Patience*: reaching out requires time and effort... (2) *Building Trust*: show genuine interest and respect... (3) *Community Protocols*: understand each community's norms... (4) *Tangible Benefits*: develop tools that directly help the community.	5	5

Table 5: English QA examples across architectures.

Model	Question / Answer	Q.Men	A.Men
Single	**Q:** 更年期的定义是什么? **A:** 更年期就是月经周期的结束，即你最后一次月经。	3	2
Dual	**Q:** 什么是脑雾（Brain fog)? **A:** 脑雾是指大脑似乎不属于你... 忘记想说的话... 注意力不集中，感觉大脑里有雾。	3	3
RAG	**Q:** 为什么更年期被认为是人生中的重要阶段? **A:** 更年期标志着最后一次月经... 意味着生育能力的结束... 对身体和心理都会产生显著影响，包括情绪波动、记忆力减退等。	4	4
Multi	**Q:** 如何从三十几岁起为应对更年期做好心理和生理上的准备? **A:** (1) 了解更年期：认识到这是女性生命中的自然阶段... (2) 健康生活方式：保持健康饮食，定期锻炼... (3) 管理压力：通过冥想、瑜伽应对情绪波动... (4) 建立支持系统：与家人朋友分享感受。	4	5

Table 6: Chinese QA examples across architectures.

with Single Agent scores of 3.96 on QA Alignment and 3.72 on Answer Fluency, it improves Hindi to 4.37 and 4.14, suggesting that agentic coordination can partially mitigate language specific errors.

To validate that low Hindi IAA reflects base model limitations rather than flaws in our evaluation framework, we replace Qwen2.5 7B with Gemma3 27B. As shown in Figure 6, IAA rises from near chance agreement to substantial agreement (0.56 to 0.95). Multi Agent consistently outperforms Single Agent under both models, confirming that agentic coordination provides genuine quality gains independent of model size.

Topic Evaluation. Using the same Multi Agent setting, we analyze topic effects in Figure 5(b). *Personal Growth* performs best overall (4.46), with

Model	Question / Answer	Q.Men	A.Men
Single	**Q:** आपके चार्टर्ड अकाउंटेंट की यात्रा की शुरुआत कैसी थी? **A:** आपकी यात्रा की शुरुआत बहुत दिलचस्प थी। [degrades into incoherent repetition]	3	1
Dual	**Q:** चार्टर्ड अकाउंटेंट की यात्रा के बारे में आपका उपाय क्या था? **A:** पहले विद्यालय में अच्छा छात्र रहा... फाउंडेशन के लिए एक निश्चित स्ट्रेटेजी अपनाई थी।	3	2
RAG	**Q:** फाउंडेशन एक्जाम के लिए किस प्रकार की प्रैक्टिस की थी? **A:** साइमल्टेनियस प्रैक्टिस किया... थियोरिकल सब्जेक्ट्स से शुरू... अकाउंटिंग स्टैंडर्ड्स और प्रोफेसर के नोट्स का उपयोग किया।	3	3
Multi	**Q:** परीक्षा खतम होने के बाद परिणाम की तैयारी के लिए समय कैसे उपयोग करें? **A:** एक नियमित प्रोग्राम बनाएं। थोड़ा आराम लें, फिर विषय-विषयी सारणी बनाएं। पिछली परीक्षाओं की जांच करें और त्रुटियों को ठीक करें। बुक्स और वीडियो का उपयोग करें।	4	5

Table 7: Hindi QA examples across architectures.

Model	Question / Answer	Q.Men	A.Men
Single	**Q:** Ce a fost un moment în care te simțeai back-stabbed? **A:** Un moment a fost când am făcut documentarea... și am reușit să facem o bună documentare.	3	2
Dual	**Q:** Cum a influențat popularitatea seriei viața ta înainte de 20 de ani? **A:** Popularitatea m-a dus la notorietate... viața era localizată înainte... cu succesul seriei m-au cunoscut mai mulți oameni.	3	3
RAG	**Q:** Cum te simți când oferi timp lucrurilor care te aduc bucurie? **A:** Mă simt realizat și plin de energie... bucuria vine din conexiunea cu oamenii... un sentiment de pură bucurie în fiecare zi.	3	4
Multi	**Q:** Cum ai merge în pregătirea pentru a face un curs de actorie? **A:** Este important să te implici profund... să nu fii satisfăcut de rezultate imediate... să te străduiești pentru calitate... dedicația și lipsa de superficialitate sunt esențiale.	4	5

Table 8: Romanian QA examples across architectures.

high *QA Alignment* (4.67) and *Answer Mentorship* (4.66). *Career & Education* (4.41) and *Business/Enterprise* (4.37) follow, while *Physical Health* scores lowest (4.09), especially on *Question Clarity* (3.75) and *Question Mentorship* (3.90), reflecting clinically dense language. *Finance* remains strong on *QA Alignment* (4.47), likely due to structured discourse. Overall, agentic coordination appears most useful for complex mentorship topics, with smaller gains for domains well represented in pretraining.

Takeaways. Our results show that: (1) task oriented metrics are reliable but more subjective than linguistic metrics; (2) LLM evaluation must be language and metric aware, motivating multi judge strategies; (3) Multi Agent models outperform simpler architectures on mentorship and alignment; and (4) agentic coordination is especially useful for complex topics and lower resource languages.

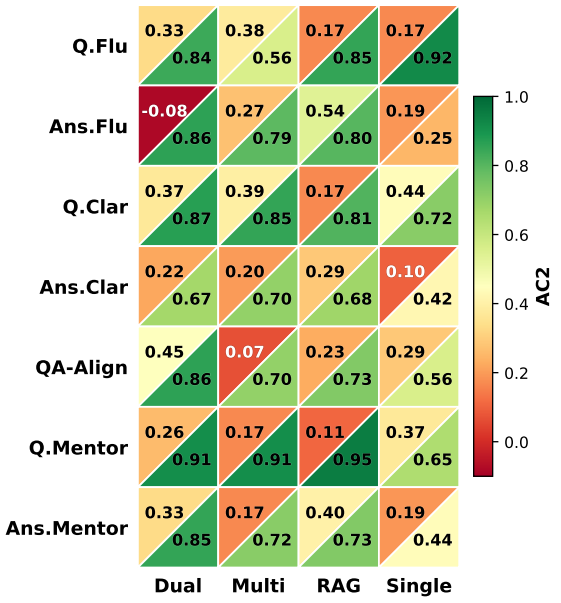


Figure 6: Hindi IAA: Qwen2.5 7B (upper ◸) vs. Gemma3 27B (lower ◿) across models and metrics.

6 Conclusion

We introduced MentorQA, the first multilingual dataset and evaluation framework for mentorship focused question answering from long form videos. By treating mentorship as a distinct QA objective, we show that existing QA benchmarks remain limited for settings that require guidance, clarity, and learning value beyond factual correctness.

Through a controlled comparison of Single Agent, Dual Agent, RAG, and Multi Agent pipelines, we find that agentic coordination consistently improves mentorship quality, especially for complex topics and lower resource languages. Building on these results, a promising future direction is a hybrid architecture that combines RAG's efficient retrieval and chunking with Multi Agent question scoring, ranking, and brainstorming, preserving quality gains while reducing cost.

Our automated evaluation analysis further shows substantial variation in alignment with human judgments across languages and metrics, highlighting the need for task aware and language sensitive evaluation. Overall, MentorQA establishes mentorship focused QA as a new research direction and provides a multilingual benchmark for studying architectures and evaluation methods that better support learning, reflection, and guidance across cultures. The dataset and models are released at https://aim-scu.github.io/MentorQA.

Limitations

Our work represents an initial step toward mentorship-focused question answering. While it provides a comprehensive evaluation of QA architectures, several limitations remain.

Generality of Findings. Our study focuses on mentorship-focused QA derived from long-form videos across four languages and a limited set of domains. This design enables controlled comparison of architectures and evaluation dimensions, but it does not capture the full diversity of mentorship contexts, interaction styles, or cultural norms found in real-world settings. In particular, our findings may not transfer to highly interactive mentorship scenarios or to languages with substantially different resource profiles. Rather than aiming for universal coverage, our goal is to establish mentorship-focused QA as a distinct problem and to provide a benchmark that supports further evaluation and analysis. We encourage future work to extend our open-source framework to additional languages and domains.

A Fixed Base Model. To isolate the effects of architectural and agentic design, we fix the underlying base language model (Qwen 2.5) across all experiments. This choice enables fair comparison but limits insight into how multi-agent pipelines interact with other foundation models. Newer multilingual or instruction-tuned models may exhibit different trade-offs or amplify agentic gains, which we invite future work to explore.

Allocation Design. To enable controlled comparison across architectures, we use a fixed question budget of $T = 20$ and a single segment strength function, $S(i) = \mu - \sigma$, throughout the Multi Agent pipeline. These choices keep the generation setting consistent, but we do not explore alternative allocation functions or question budgets in this work. Future work could examine whether different allocation strategies or values of T further improve mentorship quality or coverage.

Evaluation and Scalability. Our evaluation combines human judgments with LLM-based judges to assess mentorship quality across multiple dimensions. Although LLM judges provide scalable signal, we observe notable variation in their alignment with human judgments across languages and metrics. Accordingly, automated evaluation should be viewed as complementary rather than a substitute for human assessment. Multi-agent pipelines also incur higher computational cost than single-agent or RAG-based approaches, which may limit deployment in resource-constrained settings. Our analysis prioritizes comparative effectiveness, and we see efficiency-aware agentic designs as an important direction for future work.

Subjectivity of Mentorship Evaluation. Mentorship quality is inherently subjective, as reflected in the moderate inter-annotator agreement observed for mentorship-focused metrics. Annotators may reasonably prioritize different aspects of guidance, particularly across cultural and linguistic contexts. While our evaluation captures this variability, it does not explicitly model disagreement or multiple valid interpretations of mentorship quality. Developing evaluation protocols that better account for such diversity remains an open challenge.

Downstream Educational Impact. Our evaluation focuses on mentorship QA quality rather than downstream educational outcomes. Future work could examine how mentorship-focused QA supports comprehension, reflection, and skill development in real educational settings.

Computational Cost. On average per 1.5-hour video, Single-Agent requires 78 seconds, Dual-Agent 130 seconds, RAG 181 seconds, and Multi-Agent 316 seconds. The increased latency in Multi-Agent is primarily driven by the Question Brainstorming and Scoring steps, which involve over-generation and iterative evaluation. This overhead is justified by the substantial mentorship quality gains, particularly for lower-resource languages and complex topics where simpler architectures degrade significantly, making it a worthwhile trade-off for educational AI applications. As discussed in the Conclusion, a hybrid architecture combining RAG efficiency with Multi-Agent coordination remains a promising direction to reduce this cost.

Ethical Considerations

Mentorship content often involves personal experiences, sensitive advice, and implicit power dynamics. Although all videos used in MentorQA are publicly available, we emphasize that mentorship quality is not equivalent to factual authority, and generated QA pairs should not be treated as prescriptive guidance without human oversight. Automated systems may overgeneralize advice, miss contextual nuance, or reflect cultural biases present in training data.

We also note ethical risks in automated evaluation. LLM-based judges may encode language-

or culture-specific preferences, which can disadvantage certain communication styles or communities. Our findings underscore the continued importance of human evaluation, transparent metric design, and cautious interpretation of automatic scores, particularly in multilingual and educational settings.

We release MENTORQA as a research resource rather than a deployable mentorship system, and encourage future work to incorporate participatory evaluation, educator-in-the-loop validation, and safeguards that prioritize learner well-being.

Anonymization and Personal Information MENTORQA is constructed from publicly available mentorship videos. While the source content is public, we take steps to minimize the presence and propagation of personally identifiable information (PII) in the released dataset. We process only automatically generated transcripts and do not release raw audio or video. The released dataset contains only de-identified text and focuses on general mentorship themes, guidance, and reflective insights rather than personal narratives tied to specific individuals. We follow established best practices for ethical data use from publicly available sources and release the dataset for research purposes only.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. We also thank the members of the Inclusive and Multimodal AI (AIM) Lab at Santa Clara University for their help during the annotation process. This project was partially funded by an HPC Center Catalytic Grant and a Fletcher Jones Foundation grant.

References

Marah I. Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. Technical Report MSR-TR-2024-57, Microsoft Research.

Andrew Aquilina, Sean Diacono, Panagiotis Papapetrou, and Maria Movin. 2023. An end-to-end workflow using topic segmentation and text summarisation methods for improved podcast comprehension. *arXiv preprint*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*, volume 34.

Alessia Battisti and Sarah Ebling. 2024. Automatic annotation elaboration as feedback to sign language learners. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 46–60, St. Julians, Malta. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Mao-Siang Chen and An-Zi Yen. 2024. E-qgen: Educational lecture abstract-based question generation system. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 24), Demonstrations Track*, pages 8631–8634, Macau, China. IJCAI/AAAI.

Cohere. 2024. Command r7b. `https://cohere.com/blog/command-r7b`. Official Command R7B release.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Jing Du, Guangtao Xu, Wenhao Liu, Dibin Zhou, and Fuchang Liu. 2025. Enhancing online learning through multi-agent debates for cs university students. *Applied Sciences*, 15(11):5877.

Andrew Elaryan. 2022. Question-answering for segment retrieval on podcast transcripts. Master's thesis, University of Oklahoma.

Said Al Faraby, Adiwijaya Adiwijaya, and Ade Romadhony. 2024. Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education*, 34:1008–1045.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. QGEval: Benchmarking Multi-dimensional Evaluation for Question Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11783–11803, Miami, Florida, USA. Association for Computational Linguistics.

Google DeepMind. 2025. Gemma-3-12b-it. https://huggingface.co/google/gemma-3-12b-it. Instruction-tuned 12B Gemma 3 multimodal model.

Shasha Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. 2024. A survey on neural question generation: Methods, applications, and prospects. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 2024), Survey Track*, pages 8038–8047.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Géza Kovács. 2016. Effects of in-video quizzes on mooc lecture viewing. In *Proceedings of the Third (2016) ACM Conference on Learning@Scale*, pages 31–40, Edinburgh, UK. Association for Computing Machinery.

large-traversaal. 2025. Qwen-2.5-14b-hindi. https://huggingface.co/large-traversaal/Qwen-2.5-14B-Hindi. Large language model fine-tuned for Hindi, hosted on Hugging Face.

Lin Ma and Yuchun Ma. 2019. Automatic question generation based on mooc video subtitles and knowledge graph. In *Proceedings of the 2019 7th International Conference on Information and Education Technology (ICIET 2019)*, pages 49–53, Aizu-Wakamatsu, Japan. Association for Computing Machinery.

Wenjun Meng, Yuzhe Li, Lili Chen, and Zhaomin Dong. 2025. Using the retrieval-augmented generation to improve the question-answering system in human health risk assessment: The development and application. *Electronics*, 14(2):386.

Meta. 2024. Llama-3.2-11b-vision. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision. Multimodal vision language model with text+image inputs, released under the Llama 3.2 Community License; optimized for visual recognition, reasoning, and captioning.

Mistral AI. 2024. Ministral-8b-instruct-2410. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410. Instruction-tuned Ministral 8B model released under the Mistral Research License.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. Rquge: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada. Association for Computational Linguistics.

Antonio Moreno-Cediel, Jesús Ángel del Hoyo-Gabaldón, Eva García-López, Antonio García-Cabot, and David de Fitero-Domínguez. 2024. Evaluating the performance of multilingual models in answer extraction and question generation. *Scientific Reports*, 14:15477.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Whisper large-v3 release. https://github.com/openai/whisper. Official OpenAI Whisper large-v3.

Anu Pradhan, Alexandra Ortan, Apurv Verma, and Madhavan Seshadri. 2025. Llm-as-a-judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MEETINGQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

QwenLM Community. 2025. Improving qwen3 tokenizer efficiency for multilingual use case (issue #1400). https://github.com/QwenLM/Qwen3/issues/1400. GitHub issue discussing tokenizer inefficiency in Qwen3 for multilingual scenarios including Hindi, Italian, and German.

Sebastian Ruder and Avi Sil. 2021. Multi-domain multilingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–21, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.

Sami Saadaoui and Eduardo Alonso. 2025. Coordinated llm multi-agent systems for collaborative question-answer generation. *Knowledge-Based Systems*, 330:114627.

Mohammad Sammoudi, Ahmad Habaybeh, Huthaifa I. Ashqar, and Mohammed Elhenawy. 2025. Question-answering (qa) model for a personalized learning assistant for arabic language. In *Intelligent Systems, Blockchain, and Communication Technologies*

(ISBCom 2024), Lecture Notes in Networks and Systems, vol. 1268, pages 356–367. Springer, Cham.

Qwen Team. 2025. Qwen3 technical report.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O' Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint*.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. A practical toolkit for multilingual question and answer generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 86–94, Toronto, Canada. Association for Computational Linguistics.

Kesen Wang, Daulet Toibazar, Abdulrahman Alfulayt, Abdulaziz S. Albadawi, Ranya A. Alkahtani, Asma A. Ibrahim, Haneen A. Alhomoud, Sherif Mohamed, and Pedro J. Moreno. 2025. Multi-agent interactive question generation framework for long document understanding. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Cameron R. Wolfe. 2024. Using llms for evaluation. Substack article.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025. Qwen2.5-1m technical report. *arXiv preprint*.

Mengxia Yu, Bang Nguyen, Olivia Zino, and Meng Jiang. 2026. Context selection and rewriting for video-based educational question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 40925–40933.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Jie Zhu, Junhui Li, Yalong Wen, Xiandong Li, Lifan Guo, and Feng Chen. 2025. M³finmeeting: A multilingual, multi-sector, and multi-task financial meeting understanding evaluation dataset. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 244–266, Vienna, Austria. Association for Computational Linguistics.

A Data Collection

A.1 Video Sources

To see the full detailed source of all videos, see Table 9 and Table 10.

A.2 Topic Selection across Languages.

Topic Annotation. To ensure accurate topic categorization, each QA pair in the dataset was manually annotated and verified by human annotators, establishing ground truth topic labels for all evaluations.

Videos are selected to cover six mentorship-related topics: *Entrepreneurship, Education, Finance, Mental Health, Personal Growth*, and *Physical Health*. Topic labels are automatically assigned using *Qwen2.5-7B* (Yang et al., 2025) and subsequently verified by expert annotators with proficiency in the respective languages.

English. English videos are collected from the ACL Mentorship channel[2]. ACL Mentorship is a Year-Round Mentorship initiative of the Association for Computational Linguistics (ACL), aimed at supporting students and early-career researchers entering the field of NLP. The channel hosts virtual mentorship sessions featuring mentors from academia and industry, covering topics such as choosing a research direction, writing papers, pursuing a PhD, and maintaining work–life balance. The initiative promotes equal access to career guidance and currently supports a global community of over 2,500 members.

Romanian. Romanian videos are collected from the MindArchitect channel[3], which features conversations with coaches, educators, athletes, and other specialists. Topics span education and career development (e.g., *"How do we prepare young people for the labor market?"*), as well as personal growth and mental health (e.g., *"From perfectionism and self-criticism to the joy of play and mindful creation"*).

Chinese. The Chinese videos are collected from multiple public channels rather than a single mentorship program. We curate videos that exhibit strong mentorship characteristics across mental health, career growth, finance, and physical health. The selected content primarily consists of podcast-style conversations, discussions, and interview-based dialogues involving two or three participants, typically led by domain professionals. These discussions emphasize experience-based guidance, reflective reasoning, and practical advice drawn from real world practice. Although the videos originate from diverse creators, they are unified by

[2] https://www.youtube.com/@aclmentorship
[3] https://www.youtube.com/@MindArchitect

Model Type	Prompt
Single-Agent	System: You are an expert at educational content analysis who can identify key questions and provide clear answers. Prompt: Read the following transcript carefully and identify the 20 most important questions discussed in this content. For each question: 1. Ensure the question captures a key concept or important information from the transcript 2. Provide a clear, accurate answer to the question based only on information in the transcript 3. Make sure questions and answers cover different aspects of the content the whole transcript and don't overlap significantly 4. Select questions and answers in a balanced way from throughout the entire content, not concentrating too heavily on any single section or part 5. Answers should be in proper detail length and include only the relevant information answering the question properly with educational/mentorship value. 6. Use the same Language as of the original content. Strictly Format your response as a list of question-answer pairs, with each pair clearly marked (e.g., "Question 1:", "Answer 1:"). Strictly to make your response as structured as possible so it can be easily parsed. Also avoid any other extra words in the start and beginning and only the strict structured response. Transcript: {transcript}

Figure 7: Single-Agent prompt.

their focus on mentorship-focused discourse and sustained explanatory depth.

Hindi. Hindi videos are collected from long-form interview and podcast recordings that feature in-depth conversations with domain experts, industry practitioners, entrepreneurs, and high-performing competitive exam candidates. These interactions emphasize experiential knowledge, preparation methodologies, career decision-making, and reflective discussions on challenges and learnings. The conversational format allows speakers to elaborate on nuanced perspectives and real-world insights, making the content particularly rich for extracting mentorship-focused question–answer pairs.

B Methodology

Implementation Details. All models use the PyTorch and Hugging Face Transformers stacks, and are run in inference-only mode without any fine-tuning. Inference is executed on a GPU node equipped with four NVIDIA L40S GPUs, each with 48 GB of memory, which allows us to host the model once and parallelize calls from different videos, segments, or agents across devices.

B.1 Prompt Design

See fig. 7, fig. 8, fig. 9 and fig. 10 for the prompt design.

C Evaluation

Figure 11 shows the interface used in the human annotation process.

C.1 Can the metrics accurately assess the generated QA pairs?

C.1.1 Inter-Annotator Agreement

Figure 12 reveals substantial variation in inter-annotator agreement across languages. **Romanian exhibits the highest and most consistent agreement across metrics** $(0.49-0.81)$, demonstrating strong to near-perfect reliability in evaluating mentorship-focused quality. **Chinese displays moderate to fair agreement** $(0.27-0.47)$, with metrics like *Answer Clarity* and *QA-Alignment* achieving reasonable inter-annotator consensus. **English and Hindi show lower but still fair agreement** $(0.13-0.37$ and $0.20-0.34$ respectively), with Hindi's challenges potentially stemming from tokenization issues and weaker model support for Devanagari script, making consistent evaluation more difficult even for human annotators. Overall, these results suggest that **annotation reliability is highly language-dependent**, reinforcing the need for culturally aware evaluation frameworks when designing mentorship-focused QA assessment protocols.

C.1.2 LLM-Human Agreement

LLM Human Gwet's AC2 agreement scores per language are shown in Figure 13.

C.2 How do QA-generation models perform across metrics?

All model performance across LLM-Judges is shown in Figure 14.

Qualitative results per model are shown below:

Model Type		Prompt
Dual-Agent	Chunking Agent	System: You are an expert at analyzing transcripts and segmenting them by topic. The transcript has been split into numbered lines where each line represents a complete thought. Identify topic boundaries and assign concise topic titles. Prompt: Output a JSON list of dictionaries with these keys: - "topic": Concise descriptive title (3-7 words) - "start_line": First line number of this topic - "end_line": Last line number of this topic Rules: 1. Topics must cover consecutive line numbers 2. Entire transcript must be covered without gaps or overlaps 3. The first topic must start at line 1 4. The last topic must end at line {total_lines} 5. Use line numbers exactly as provided 6. Output ONLY the JSON with no additional text Numbered Transcript: {numbered_transcript}
	QA Agent	System: You are an expert at educational content analysis who can identify key questions and provide clear answers. Prompt: Read the following transcript carefully and identify the {num_questions} most important questions discussed in this content. For each question: 1. Ensure the question captures a key concept or important information from the transcript chunk 2. Provide a clear, accurate answer to the question based only on information in this chunk 3. Make sure questions and answers cover different aspects of the content the whole chunk and don't overlap significantly 4. Select questions and answers in a balanced way from throughout the entire content, not concentrating too heavily on any single section. 5. Answers should be in proper detail length and include only the relevant information answering the question properly with educational/mentorship value. 6. Use the same Language as of the original content. Strictly Format your response as a list of question-answer pairs, with each pair clearly marked (e.g., "Question 1:", "Answer 1:"). Strictly to make your response as structured as possible so it can be easily parsed. Also avoid any other extra words in the start and beginning and only the strict structured response. Transcript: {transcript}

Figure 8: Dual-Agent prompt.

Single-Agent **Q:** What is the most important criterion when reviewing PhD applications?
A: Passion is the most important criterion when reviewing PhD applications.
Scores: QFlu 4, AFlu 5, QClar 5, AClar 1, QA-Algn 5, QMen 4, AMen 1

Dual-Agent **Q:** Why is it important to interact with students according to the speaker in the transcript?
A: Interacting with students is one of the most rewarding aspects of the academic path, as it allows the speaker to work with brilliant PhD students and other students across various levels, including undergraduates, masters, and postdoctoral fellows.
Scores: QFlu 5, AFlu 5, QClar 4, AClar 4, QA-Algn 5, QMen 4, AMen 3

RAG **Q:** How do PhD admission processes differ internationally?
A: PhD admission processes vary internationally in terms of interview conduct and committee structures. For instance, MIT has a stringent committee system where individual professors' admissions decisions are heavily scrutinized, whereas Stanford uses a rolling admissions process. Universities like UIUC or Hong Kong allow faculty members significant autonomy in admitting PhD students. This diversity highlights the importance of considering a broader range of institutions beyond prestigious American universities.
Scores: QFlu 5, AFlu 5, QClar 4, AClar 3, QA-Algn 5, QMen 5, AMen 3

Multi-Agent **Q:** What differences exist between PhD admission processes in various academic in-

Model Type		Prompt
RAG	Query-1	System: You are an expert content analyst. Your task is to read a long transcript and identify potential questions that cover the most important educational and mentorship-related topics discussed. Prompt: Based on the following transcript, generate a list of exactly 20 diverse and high-value questions. Guidelines for questions: - Focus on key concepts, advice, and actionable insights. - Ensure questions span the entire transcript, from beginning to end. - Avoid trivial or overly specific questions. - Phrase them as clear, standalone questions. Format your output STRICTLY as a numbered list. Do not add any other text before or after the list. Transcript: {transcript}
	Query-2	System: You are an expert Q&A agent. You will be given a question and a set of context paragraphs. Your task is to synthesize a clear and accurate answer based ONLY on the retrieved context. Prompt:Please answer the following question using ONLY the information from the context provided below. Question: {question} Context: {context} If the context does not contain the answer, state that the information is not available in the provided context.

Figure 9: RAG prompt.

stitutions?

A: "The differences between PhD admission processes in various academic institutions primarily revolve around the structure of the selection committees, the role of individual faculty members, and the criteria used for evaluating applications. For instance:

1. **Committee Structure**: At some institutions, like MIT, there is a strict committee process where individual professors cannot independently admit students without consensus from the committee. This contrasts with universities like UIUC or Hong Kong, where faculty members have more autonomy to select students based on their own preferences.

2. **Faculty Autonomy**: In some systems, such as those in the U.S., faculty members play a significant role in admitting PhD students. They may have the authority to interview and accept candidates directly, which can vary widely depending on the institution and department.

3. **Application Review Process**: The number of applications reviewed by each committee member differs. Some may skim through a large volume of applications, while others may focus on specific candidates they have been contacted by. For example, one panelist mentioned receiving numerous unsolicited emails, indicating that some professors may not read every application thoroughly.

4. **Criteria for Selection**: While the core process involves evaluating academic records, letters of recommendation, and research proposals, the emphasis on these factors can differ. For example, a close match in research interests might be more beneficial for some institutions, especially smaller programs, compared to larger committees that consider a broader range of criteria.

Understanding these differences can help PhD aspirants tailor their applications and outreach strategies to align better with the specific requirements and preferences of the institutions they are targeting."

Scores: QFlu 5, AFlu 5, QClar 5, AClar 5, QA-Algn 5, QMen 5, AMen 5

C.3 How does QA-generation performance vary across languages and mentorship topics?

Language Evaluation. As shown in Figure 15, English demonstrates the strongest performance across all metrics, which is expected given its prevalence in training data. Romanian and Chinese follow closely, performing comparably to English on critical dimensions like QA-Alignment and Mentorship metrics. However, Hindi exhibits a notable performance drop across all evaluated metrics. This disparity can be attributed to ar-

Model Type		Prompt
Multi-Agent	Chunking Agent	System: You are an expert at analyzing transcripts and segmenting them by topic. The transcript has been split into numbered lines where each line represents a complete thought. Identify topic boundaries and assign concise topic titles. Prompt: Output a JSON list of dictionaries with these keys: - "topic": Concise descriptive title (3-7 words) - "start_line": First line number of this topic - "end_line": Last line number of this topic Rules: 1. Topics must cover consecutive line numbers 2. Entire transcript must be covered without gaps or overlaps 3. The first topic must start at line 1 4. The last topic must end at line {total_lines} 5. Use line numbers exactly as provided 6. Output ONLY the JSON with no additional text Numbered Transcript: {numbered_transcript}
	Question Brainstorming Agent	System: You are a curious and insightful analyst. Prompt: Based on the following text, generate a list of potential questions with high educational or mentorship value. Format your output as a simple numbered list. Avoid generating duplicate questions with similar meanings. Text Segment: {segment_content}
	Scoring & Justifier Agent	System: You are an expert content evaluator. Return ONLY a number 1-10. Prompt: Rate the following question for educational/mentorship value (1=poor, 10=excellent). Question: {question} Hint: {context_hint}. Return just a number. System: You are a content analyst. Your task is to provide a clear and concise justification for a question's selection status. Prompt: You will be given a question, its source topic, and its "Selected" or "Rejected" status. Your job is to explain *why* that status makes sense. **Question:** {question_item['question']} **Source Topic:** {question_item['source_segment']['topic']} **Selection Status:** {question_item['status']} Based on the information above, provide a concise reason. Use the following as **sample reasons for style and inspiration, but you are not limited to them**: - **Good `Selected` reasons:** - "Addresses the key concept of 'model gluing', providing high educational value." - "Asks about career transition challenges, which has direct mentorship value." - **Good `Rejected` reasons:** - "The question is too basic and offers little insight." - "This topic is already covered by another, more specific selected question." - "This question has a low score based on the result of agent 3."
	Answering Agent	System: You are a helpful Q&A assistant. Your task is to answer the given question using ONLY the information from the provided context. Prompt: **Context:** {context} — **Question:** {question} Based on the context above, what is the answer of the following questions ?

Figure 10: Multi-Agent prompt.

chitectural and data constraints in the Qwen2.5 model. The model employs a Byte-Level Byte Pair Encoding (BBPE) tokenizer that is specifically optimized for English and Chinese (Yang et al., 2025). When processing Hindi text in Devanagari script, the tokenizer fails to recognize complete word units and instead fragments individual words into numerous byte-level tokens. This suboptimal tokenization, combined with limited Hindi representation in the training corpus, directly impacts the model's ability (QwenLM Community, 2025) to generate and evaluate Hindi content effectively. This interpretation is further supported by the model's reported MMLU benchmark scores: English achieves 74.37 while Hindi scores only 52.16 (large-traversaal, 2025), reflecting a substantial performance gap. The combination of tokenization inefficiency and insufficient training exposure explains the observed quality degradation in Hindi-generated QA pairs.

The Multi-Agent model demonstrates a consistent performance advantage, effectively raising the ceiling for every language compared to the other three architectures. This improvement is evident across both high- and low-resource languages. A particularly striking improvement is observed for Hindi, which reveals the Multi-Agent model's role as a performance equalizer. In the Single-Agent architecture, Hindi substantially underperforms compared to Chinese, English, and Romanian, with QA-Alignment dropping to 3.96 and Answer Fluency falling to 3.72—both failing

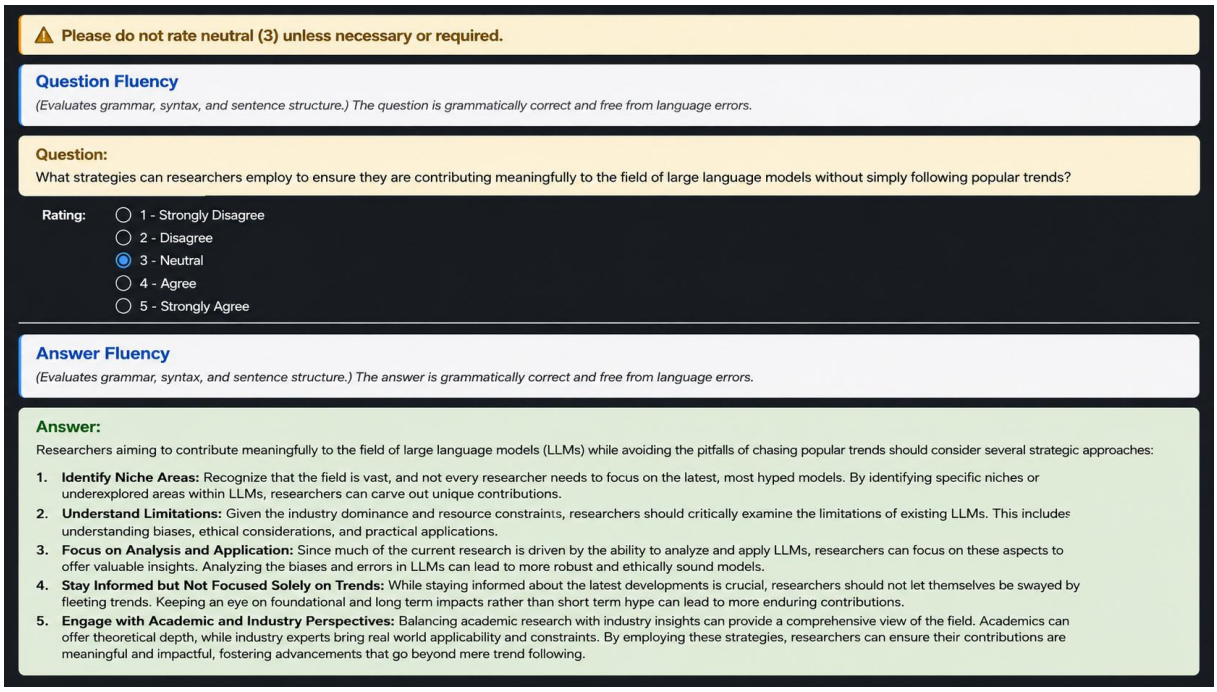


Figure 11: Human evaluation interface used to rate question and answer quality on the 5 point Likert scale.

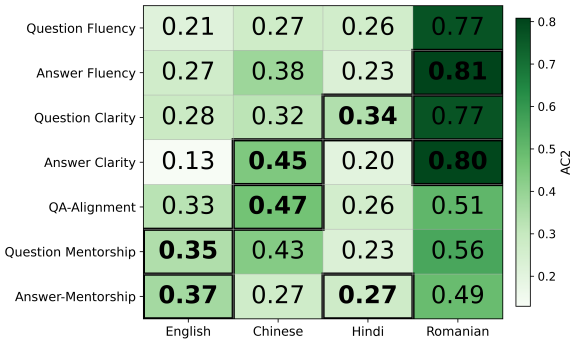


Figure 12: Inter-Annotator Agreement AC2 Scores by Language

to reach the 4.0 threshold. However, the Multi-Agent model significantly boosts these scores, elevating QA-Alignment to 4.37 and Answer Fluency to 4.14. This finding suggests that even when a base LLM exhibits inherent weaknesses in a specific language, implementing a Multi-Agent architecture can effectively compensate for these limitations. The approach acts as a performance equalizer, enabling the model to deliver quality that approaches or matches that of high-resource languages, thereby bypassing the model's native capability constraints.

Topic Evaluation. As illustrated in Figure 16, topic complexity substantially impacts LLM judge performance across content domains. To ensure accurate topic categorization, each QA pair in the dataset was manually annotated and verified by human annotators, establishing ground truth topic labels for all evaluations. **Personal Growth.** This domain achieves the highest scores in both Mentorship (4.55) and Alignment (4.56). The prevalence of self-help content and motivational literature in pre-training corpora enables models to generate well-structured guidance with relative ease, as the domain relies on broadly applicable principles rather than specialized knowledge. **Physical Health.** This domain presents the greatest challenge, registering the lowest scores in Mentorship (4.04) and Alignment (4.12). Qualitative analysis reveals that speakers frequently provided medical-style guidance incorporating complex anatomical terminology (e.g., hippocampus, amygdala, prefrontal cortex), biochemical nomenclature (e.g., cortisol, dopamine, serotonin), bodily fluids, and pharmaceutical references. Models demonstrated insufficient precision for assessing such nuanced clinical content, frequently producing vague or overly generalized evaluations. **Finance.** Despite having the smallest dataset (287 samples), Finance achieved a robust Alignment score of 4.47, comparable to more straightforward topics. This success stems from the structured, rule-based nature of financial discourse. Speakers typically provided logically structured guidance adhering to consistent frameworks (e.g., profit-loss analysis, investment strategies), enabling accurate LLM evalua-

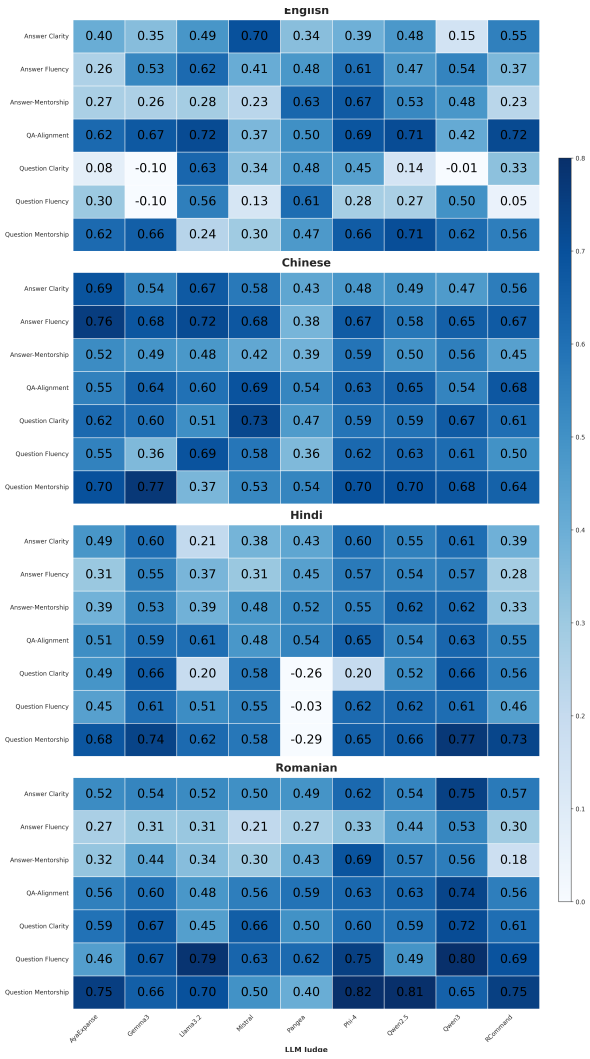


Figure 13: LLM-Human AC2 Agreement Scores

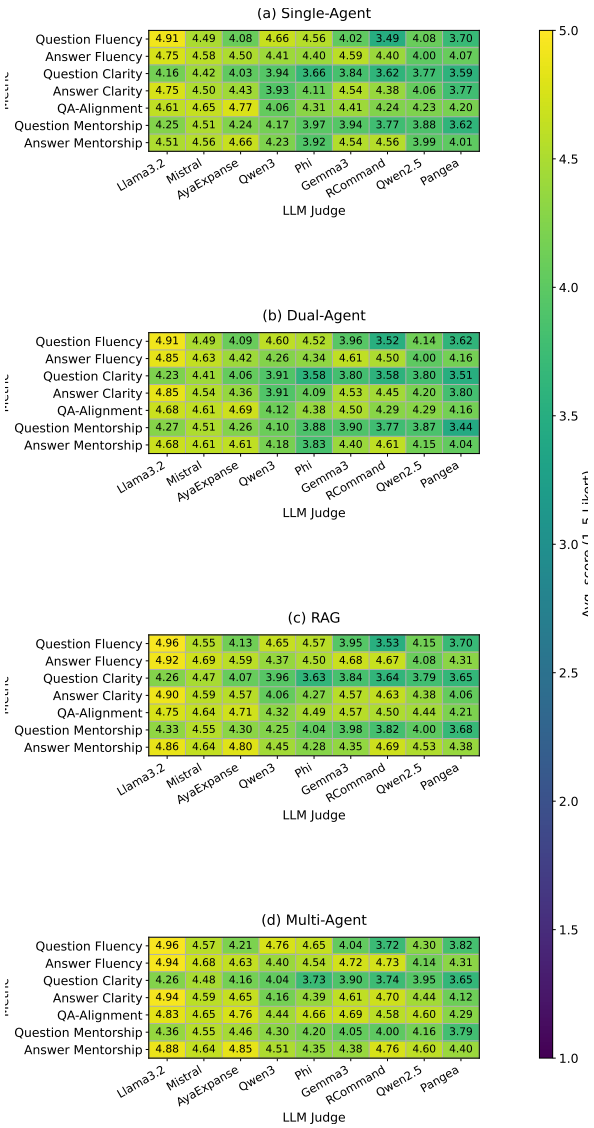


Figure 14: Judge-wise performance across QA-generation models: (a) Single-Agent, (b) Dual-Agent, (c) RAG, (d) Multi-Agent. Scores are averaged over metrics using a shared 1–5 Likert scale.

tion even with limited training examples.

As demonstrated in Figure 16, architectural complexity significantly impacts performance on challenging domains while showing minimal differentiation on straightforward topics. **Physical Health.** This domain exhibits the most pronounced performance gap. Single Agent achieves Answer Mentorship of 3.78 and QA Alignment of 3.95, with marginal improvement in Dual Agent (3.94 and 4.01). Multi Agent substantially elevates these to 4.24 and 4.34, demonstrating clear advantages in handling clinically complex content. **Finance.** This logic intensive domain reveals similar architectural differentiation. Single Agent produces Answer Mentorship of 4.21, Q Mentorship of 3.86, QA Alignment of 4.23, and Answer Clarity of 4.23. Multi Agent markedly improves these to 4.51, 4.20, 4.60, and 4.50, respectively. **Personal Growth and Career.** These domains show minimal performance variation across architectures, with Single Agent achieving scores comparable to Multi Agent. This reflects the abundance of motivational content in training data, enabling simpler architectures to generate effective responses. These results reveal domain dependent performance differentiation: common topics (Career, Personal Growth) show architectural equivalence, while challenging domains (Physical Health, Finance) demonstrate significant separation—Single and Dual Agent models degrade substantially, whereas Multi Agent maintains robust performance.

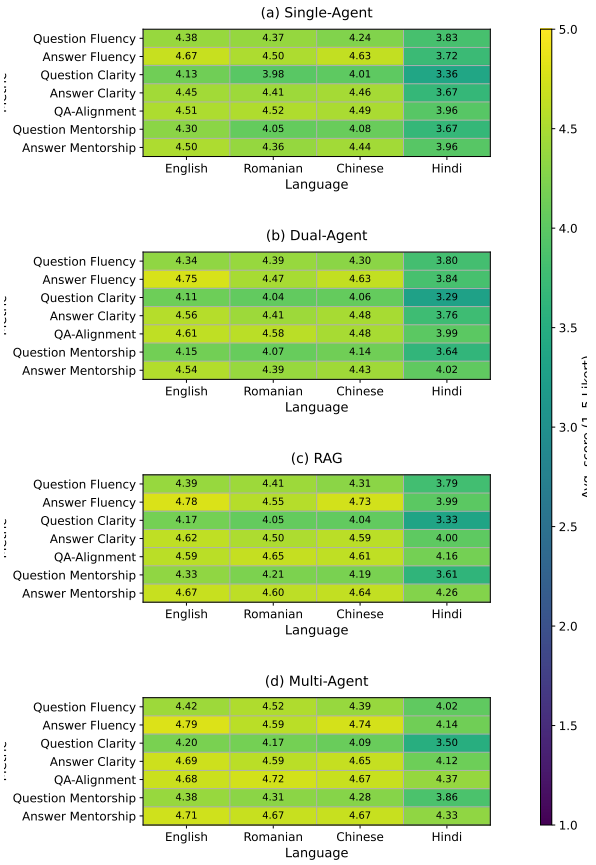


Figure 15: Language-wise performance across QA-generation models: (a) Single-Agent, (b) Dual-Agent, (c) RAG, (d) Multi-Agent. Scores are averaged over human and LLM judges (1–5 Likert).

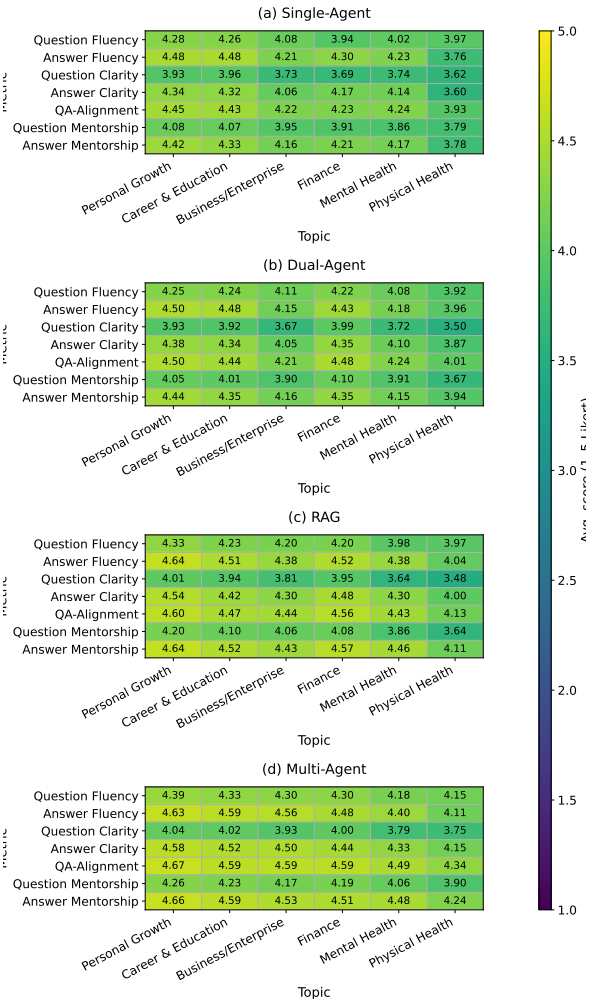


Figure 16: Topic-wise performance across QA-generation models: (a) Single-Agent, (b) Dual-Agent, (c) RAG, and (d) Multi-Agent. Scores are averaged over human and LLM judges (1–5 Likert scale).

Table 9: Mentorship Videos

Video ID	URL	Language	Video ID	URL	Language
1	Video Link	English	31	Video Link	Chinese
2	Video Link	English	32	Video Link	Chinese
3	Video Link	English	33	Video Link	Chinese
4	Video Link	English	34	Video Link	Chinese
5	Video Link	English	35	Video Link	Chinese
6	Video Link	English	36	Video Link	Chinese
7	Video Link	English	37	Video Link	Chinese
8	Video Link	English	38	Video Link	Chinese
9	Video Link	English	39	Video Link	Chinese
10	Video Link	English	40	Video Link	Chinese
11	Video Link	English	41	Video Link	Chinese
12	Video Link	English	42	Video Link	Chinese
13	Video Link	English	43	Video Link	Chinese
14	Video Link	English	44	Video Link	Chinese
15	Video Link	English	45	Video Link	Chinese
16	Video Link	English	46	Video Link	Chinese
17	Video Link	English	47	Video Link	Chinese
18	Video Link	English	48	Video Link	Chinese
19	Video Link	English	49	Video Link	Chinese
20	Video Link	English	50	Video Link	Chinese
21	Video Link	English	51	Video Link	Chinese
22	Video Link	English	52	Video Link	Chinese
23	Video Link	English	53	Video Link	Chinese
24	Video Link	English	54	Video Link	Chinese
25	Video Link	English	55	Video Link	Chinese
26	Video Link	English	56	Video Link	Chinese
27	Video Link	English	57	Video Link	Chinese
28	Video Link	English	58	Video Link	Chinese
29	Video Link	English	59	Video Link	Chinese
30	Video Link	English	60	Video Link	Chinese

Table 10: Mentorship Videos (Continued)

Video ID	URL	Language	Video ID	URL	Language
61	Video Link	Hindi	91	Video Link	Romanian
62	Video Link	Hindi	92	Video Link	Romanian
63	Video Link	Hindi	93	Video Link	Romanian
64	Video Link	Hindi	94	Video Link	Romanian
65	Video Link	Hindi	95	Video Link	Romanian
66	Video Link	Hindi	96	Video Link	Romanian
67	Video Link	Hindi	97	Video Link	Romanian
68	Video Link	Hindi	98	Video Link	Romanian
69	Video Link	Hindi	99	Video Link	Romanian
70	Video Link	Hindi	100	Video Link	Romanian
71	Video Link	Hindi	101	Video Link	Romanian
72	Video Link	Hindi	102	Video Link	Romanian
73	Video Link	Hindi	103	Video Link	Romanian
74	Video Link	Hindi	104	Video Link	Romanian
75	Video Link	Hindi	105	Video Link	Romanian
76	Video Link	Hindi	106	Video Link	Romanian
77	Video Link	Hindi	107	Video Link	Romanian
78	Video Link	Hindi	108	Video Link	Romanian
79	Video Link	Hindi	109	Video Link	Romanian
80	Video Link	Hindi	110	Video Link	Romanian
81	Video Link	Hindi	111	Video Link	Romanian
82	Video Link	Hindi	112	Video Link	Romanian
83	Video Link	Hindi	113	Video Link	Romanian
84	Video Link	Hindi	114	Video Link	Romanian
85	Video Link	Hindi	115	Video Link	Romanian
86	Video Link	Hindi	116	Video Link	Romanian
87	Video Link	Hindi	117	Video Link	Romanian
88	Video Link	Hindi	118	Video Link	Romanian
89	Video Link	Hindi	119	Video Link	Romanian
90	Video Link	Hindi	120	Video Link	Romanian