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

Anonymous ACL submission

Abstract

Question answering systems are typically evaluated on factual correctness, yet many real-world applications—such as education and career guidance—require *mentorship*: responses that provide reflection and guidance. Existing QA benchmarks rarely capture this distinction, particularly in multilingual and long-form settings. We introduce MentorQA, the first multilingual dataset and evaluation framework for mentorship-focused question answering from long-form videos, comprising nearly 9,000 QA pairs from 180 hours of content across four languages. We define mentorship-focused evaluation dimensions that go beyond factual accuracy, capturing clarity, alignment, and learning value. Using MentorQA, we compare Single-Agent, Dual-Agent, RAG, and Multi-Agent QA architectures under controlled conditions. Multi-Agent pipelines consistently produce higher-quality mentorship responses, with especially strong gains for complex topics and lower-resource languages. We further analyze the reliability of automated LLM-based evaluation, observing substantial variation in alignment with human judgments. Overall, this work establishes mentorship-focused QA as a distinct research problem and provides a multilingual benchmark for studying agentic architectures and evaluation design in educational AI. The dataset and evaluation framework are released at https://anonymous.4open.science/r/MentorQA/.

1 Introduction

Question answering (QA) systems are typically evaluated on factual correctness (Mulla and Gharpure, 2023; Fu et al., 2024). However, in many real-world settings—such as education, career guidance, and personal development—users also seek *mentorship*: answers that provide context, reflection, and guidance rather than isolated facts (Ma and Ma, 2019; Kovács, 2016). Existing QA benchmarks and models rarely capture this distinction, particularly in multilingual and long-form scenarios. This gap is especially evident for long-form mentorship content, where relevant guidance is sparsely distributed across ex-

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), Mentorship (Q/AMen) on 1-5 Likert scale.

tended narratives. Answering mentorship-focused questions requires identifying and synthesizing insights that support learning and decision-making, going beyond surface-level correctness. While recent language models have advanced open-domain QA (Guo et al., 2024), it remains unclear how well they support mentorship-focused question answering across languages with different discourse norms and resource availability.

Prior QA research has largely focused on short-form inputs, factual accuracy, and monolingual evaluation (Faraby et al., 2024). Multilingual QA benchmarks emphasize linguistic fluency and correctness, but seldom assess whether responses provide meaningful guidance or learning value. As a result, datasets, metrics, and systematic evaluations for *mentorship-focused multilingual QA from long-form content* are largely missing. Moreover, the impact of architectural choices—such as retrieval augmentation or agentic coordination—on mentorship quality remains underexplored.

Contributions. **First, we introduce MentorQA, the first multilingual dataset and evaluation framework designed for mentorship-focused question answering from long-form videos.** The dataset contains nearly 9,000 QA pairs derived from 180 hours of mentorship content across four languages. **Second, we define mentorship-focused evaluation dimensions that capture clarity, alignment, and guidance, and compare four 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. **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 has explored extracting questions and answers from long educational recordings to support learning. Existing systems generate questions from lecture summaries (Chen and Yen, 2024), select or rewrite context for video-based educational QA (Yu et al., 2025), or apply segment-level retrieval and extractive QA to podcasts and meeting transcripts (Elaryan, 2022; Prasad et al., 2023). Other efforts move toward personalization, including QA for language learning assistants (Sammoudi et al., 2025) and multi-agent debate frameworks for online education (Du et al., 2025). While these approaches address educational content, they primarily optimize for factual correctness or content coverage and are often monolingual or tied to a single modeling paradigm. In contrast, mentorship-focused QA emphasizes guidance, reflection, and learning value—objectives that remain underexplored in long-form, multilingual settings. Our work targets this gap by extracting mentorship-focused QA from long-form videos across multiple languages and evaluating architectural choices under a unified framework.

Agentic and RAG-Based QA Pipelines. Agentic models have been proposed to support complex, multi-step workflows for long-document understanding and QA. Prior work demonstrates how specialized agents can coordinate to decompose reading, reasoning, and generation tasks (Wang et al., 2025; Saadaoui et al., 2025), and recent surveys analyze common collaboration mechanisms in such systems (Tran et al., 2025). Retrieval-augmented generation (RAG) has also been widely adopted for domain-specific QA, including applications in health and risk assessment (Meng et al., 2025), as well as multi-stage pipelines for podcast and meeting analysis (Aquilina et al., 2023; Zhu et al., 2025). However, these approaches are typically evaluated on generic QA objectives or task efficiency, and rarely compare architectural paradigms under controlled conditions. Our work differs in two key ways: we systematically compare Single-Agent, Dual-Agent, RAG, and Multi-Agent pipelines using a fixed base model, and we evaluate them with respect to mentorship-focused quality rather than factual accuracy alone, revealing when agentic coordination is most beneficial.

Multilingual QA and Metric-Based Evaluation. A parallel body of work has examined evaluation methods for question generation and the extension of QA beyond English. Early studies investigated automatic QG metrics (Nema and Khapra, 2018), followed by reference-free and answer-based approaches such as RQUGE (Mohammadshahi et al., 2023) and QG-Eval (Fu et al., 2024), as well as LLM-as-a-judge frameworks for rubric-driven

evaluation (Wolfe, 2024). Multilingual QA benchmarks further explore QA and question generation across languages and domains (Ushio et al., 2023; Moreno-Cediel et al., 2024; Ruder and Sil, 2021; Asai et al., 2021). However, recent analyses show that many automatic metrics correlate weakly with human judgments, particularly for task-oriented dimensions such as usefulness or guidance (Fu et al., 2024). Motivated by these limitations, we introduce a streamlined evaluation framework with mentorship-focused metrics designed to better capture learning value and to enable systematic comparison between human and automated evaluation in multilingual QA.

3 The MentorQA Dataset

We introduce MentorQA, to our knowledge the first QA dataset focused on mentorship. Unlike traditional QA benchmarks that emphasize factual recall, 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 consists of 8,990 QA pairs drawn from 120 long-form mentorship videos, covering four languages and six mentorship topics. We next describe our data collection pipeline, including video selection and QA generation.

Video Collection and Processing. We collect long-form mentorship videos from YouTube, focusing on podcast-style conversations and panel discussions that provide reflective, experience-driven guidance. Each video averages over 1.5 hours, resulting in more than 100 hours of multilingual mentorship content. We process only the audio track, transcribing it with Whisper[1], which automatically detects the source language and produces transcripts used as input to the QA models.

Languages. The dataset covers four languages—*English*, *Chinese*, *Hindi*, and *Romanian*—selected to span typologically and culturally diverse settings while supporting reliable, high-quality annotation. All videos and QA pairs are balanced across languages and collected and evaluated by native or expert speakers, ensuring culturally informed analysis and evaluation.

Topic Selection across Languages. Videos are selected to cover six mentorship-related topics[2]: *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.

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.

English videos focus on academic and professional mentorship, including research directions, publishing, PhD pathways through panel discussion with experts in academia and industry, such as ACL Mentorship.[3]

Romanian videos center on education, personal development, and mental health through conversations with professionals from diverse backgrounds, such as authors, coaches or athletes.[4]

Chinese videos center on experience-driven mentorship in areas such as mental health, career growth, finance, and physical well-being.[5]

Hindi videos focus on practical guidance related to career advancement, business, and mental or physical health.[6]

QA Generation. We generate question–answer pairs from each video transcript using four complementary QA-generation models: *Single-Agent*, *Dual-Agent*, *RAG*, and *Multi-Agent*. These models are designed to address challenges of long-form, conversational mentorship content, including long-context reasoning and topic drift. For each video, each model generates approximately 20 mentorship-focused QA pairs, yielding about

[1] https://github.com/openai/whisper
[2] All the mentorship videos are in Appendix B.
[3] https://www.youtube.com/@aclmentorship
[4] https://www.youtube.com/@MindArchitect
[5] https://www.youtube.com/@DrinkingLibrary
[6] https://www.youtube.com/@ranveerallahbadia

80 QA pairs per video. Across 120 videos, this process results in 8,990 QA pairs (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* and *Clarity* (linguistic quality), and *QA Alignment* and *Mentorship Value* (task-oriented quality). Human evaluation is conducted by three annotators per language, complemented by nine LLM-based judges to support scalable analysis and study alignment with human judgments. Evaluation is described in Section 5.

PII Anonymization. Although the source data are public, we proactively reduce the presence and downstream propagation of personally identifiable information (PII). We apply a layered anonymization strategy that combines automatic detection—using named entity recognition based on XLM-RoBERTa[7], structured PII detection and anonymization via Presidio[8], and LLM-based prompts to flag implicit identifiers—with human review, which is especially important for lower-resource languages and culturally specific entities.

Scalability. MentorQA is released as an extensible, open-source resource. While the current version covers four languages, a limited number of channels, and a single video platform, the codebase and documentation support direct expansion to additional languages and content sources.

4 Mentorship-focused QA Generation

Mentorship-focused question answering from long-form conversational videos poses challenges that go beyond standard QA generation. In addition to long-context reasoning and topic shifts, mentorship content exhibits a critical property: *mentorship value is unevenly distributed*. Short segments rich in reflection, advice, or experience-driven insight often carry more educational value than longer descriptive or narrative portions. Effective mentorship-focused QA must therefore identify, prioritize, and allocate questions based on the *quality* of guidance rather than transcript length alone.

To study this problem systematically, we design four QA-generation pipelines of increasing structure: *Single-Agent*, *Dual-Agent*, *Retrieval-Augmented Generation (RAG)*, and *Multi-Agent*. These pipelines progressively introduce (1) topic awareness, (2) answer grounding, and (3) explicit control over mentorship quality and question allocation.

Across all pipelines, we use the same base language model, *Qwen-2.5 7B-Instruct-1M* (Yang et al., 2025), configured in long-context mode. This controlled setup ensures that performance differences arise from architectural choices rather than model capacity. Full prompts are provided in subsection C.1.

4.1 Single-Agent

The Single-Agent baseline processes the full video transcript using a single long-context LLM. The model is instructed via a system prompt to act as an expert educational analyst and to generate mentorship-focused questions and answers. The prompt enforces broad topical coverage, discourages trivial or localized questions, and requires answers that provide educational or mentorship value. For each transcript, the model generates a fixed set of 20 question–answer pairs spanning the entire conversation. This baseline relies solely on the long-context capabilities of the underlying LLM, without explicit segmentation, retrieval, or coordination mechanisms.

4.2 Dual-Agent

While the Single-Agent baseline can generate high-quality questions, it often over-concentrates QA pairs in dense or early transcript regions. To address this, the Dual-Agent pipeline separates topic segmentation from QA generation.

A *Chunking Agent* first partitions the transcript into coherent, topic-based segments. The agent is instructed to cover the full conversation while avoiding overly small or fragmented segments. Segment lengths adapt dynamically to topic duration, and the content itself remains unchanged.

Each segment is then passed to a *QA Agent*, which generates a fixed number of mentorship-focused QA pairs per segment using the same prompt as the Single-Agent baseline. This topic-aware decomposition yields more balanced coverage across the conversation and improves alignment with underlying mentorship themes.

4.3 Retrieval Augmented Generation (RAG)

RAG explicitly separates question generation from answer grounding. In this pipeline, the model

[7] https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl

[8] https://github.com/microsoft/presidio

first generates a fixed set of diverse, mentorship-focused questions from the full transcript. Each question is then answered using only relevant transcript segments retrieved via similarity search. We construct a multilingual retrieval index using the BGE-M3 embedding model (Chen et al., 2024), which automatically chunks long transcripts and supports all four dataset languages. For each question, the most relevant transcript chunks are retrieved and provided as context to a second LLM query, which generates the final answer. This approach enables global question coverage while keeping answers explicitly tied to relevant transcript segments.

4.4 Multi-Agent

While the Dual-Agent model improves coverage through topic-aware segmentation and RAG grounds answers via retrieval, both approaches rely on a single-pass question generation strategy. *In practice, we observe substantial variation in mentorship quality across topics: uniformly allocating questions can over-represent weak segments and under-represent parts of the conversation rich in guidance.* To address this, we introduce a Multi-Agent framework that explicitly assesses mentorship quality and adaptively allocates questions across topics based on their estimated value. As shown in Figure 1, the pipeline decomposes QA generation into specialized agents responsible for topic discovery, question exploration, quality assessment, allocation, and answer generation.

The process begins with the same *Chunking Agent* used in the Dual-Agent setting, which performs dynamic topic-based segmentation and assigns each segment a concise title. For each segment, a *Question Brainstorming Agent* deliberately over-generates candidate questions, emphasizing diversity and mentorship relevance. A *Scorer Agent* then assigns each candidate question q an integer mentorship score $score(q) \in \{1, \ldots, 10\}$. To capture segment-level importance, we compute a *segment strength* score:

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

where $\mu(q)$ and $\sigma(q)$ denote the mean and standard deviation of question scores within segment i. This formulation favors segments with consistently high mentorship value while penalizing segments with noisy or uneven question quality. Given a target of T questions per video, each segment receives a quota proportional to its strength:

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

with a minimum of one question per non-empty segment. Within each segment, only the top-ranked questions are retained. To improve transparency and reduce redundancy, a *Justifier Agent* generates short explanations for question selection or rejection. Finally, an *Answering Agent* produces the final answers, grounding each response in its corresponding segment and emphasizing mentorship-oriented explanations.

Overall, this design enables broader exploration of the question space, explicit assessment of mentorship quality, and adaptive allocation of questions to the most valuable parts of long-form mentorship conversations.

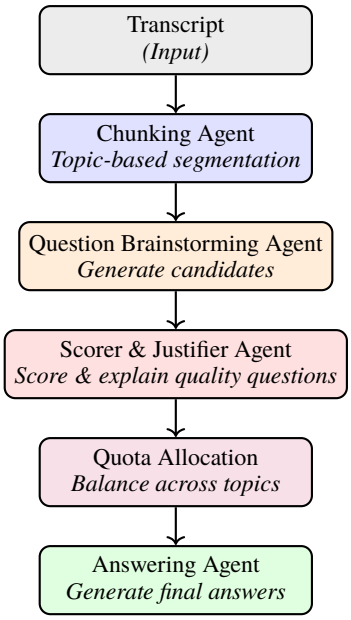

Figure 1: Multi-Agent QA generation pipeline. The framework decomposes QA generation from long-form transcripts into specialized agents for topic segmentation, question exploration, quality scoring, quota-based allocation, and grounded answer generation. This design enables explicit control over mentorship quality and balanced question allocation across topics.

5 Evaluation Methodology

We evaluate the QA-generation models using both human and automatic evaluation. Beyond standard linguistic and alignment measures, we introduce mentorship-focused evaluation dimensions that capture guidance, reflection, and learning value—properties not addressed by current QA evaluation frameworks.

5.1 Evaluation Metrics

Evaluating mentorship-focused QA requires going beyond factual correctness. We build on QG-Eval (Fu et al., 2024), which distinguishes linguistic and task-oriented dimensions, and extend it to explicitly capture alignment and mentorship value. While Fluency and Clarity are adopted from prior work, we evaluate them separately for questions and answers to enable more fine-grained analysis.

Linguistic Metrics. We assess surface-level language quality using the following metrics, applied independently to questions and answers:

- **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. While linguistic quality is a prerequisite for usable QA, our primary focus is on task-oriented metrics that capture alignment and mentorship effectiveness.

- **QA Alignment**: *The answer directly corresponds to what the question asks. The question–answer pair demonstrates proper alignment, where the answer adequately satisfies the question.*
- **Question Mentorship**: *The question provides 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—dimensions not addressed by existing QA evaluation frameworks.

Rating Scale. All metrics are rated on a 5-point Likert scale (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 distinctions in metric quality and resulted in lower inter-annotator agreement.

5.2 Human Evaluation

We recruit 12 annotators, organized into four language-specific groups corresponding to the dataset languages, with three native or fluent speakers per language. Annotators evaluate a balanced sample of QA pairs drawn from five videos per language and evenly distributed across all four QA-generation models. This results in 240 unique QA pairs, each receiving three independent ratings. An example of the annotation interface is provided in Appendix D.

Inter-Annotator Agreement. We measure annotation agreement using Gwet's AC2, computed separately for each evaluation metric across languages and models. We choose Gwet's AC2 over Krippendorff's alpha, which has been shown to produce misleading agreement estimates in skewed rating distributions common in AI system evaluation (Battisti and Ebling, 2024; Pradhan et al., 2025).

5.3 Automatic Evaluation

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

We select nine open-source or publicly documented multilingual models to ensure reproducibility: Qwen2.5-7B (Yang et al., 2025), Command-R-7B (Cohere and Aakanksha, 2025), Mistral-8B-Instruct (Mistral AI, 2025), Llama3.2-11B (Meta, 2024), Qwen3-8B (Team, 2025), Gemma3-12B-IT (Google, 2025), Pangea-7B (Yue et al., 2024), AyaExpanse-8B (Dang, 2024), and Phi-4-15B (Microsoft, 2025).

LLM-Human Agreement. We assess LLM judge effectiveness by computing Gwet's AC2 between LLM-based scores and human annotations for each evaluation metric, across languages and QA-generation models. This analysis allows us to quantify when automatic evaluation aligns with human judgments in mentorship-focused QA.

6 Evaluation Results

We evaluate mentorship-focused QA generation along four complementary axes: (1) the reliability of mentorship-focused evaluation, (2) the scalability of mentorship-focused evaluation, (3) the effectiveness of different QA-generation models, and (4) the consistency of these effects across languages and topics.

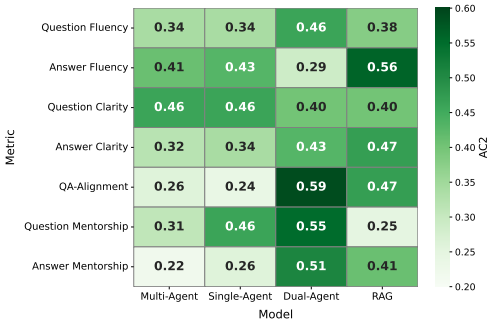

Figure 2: Inter-Annotator Agreement Scores.

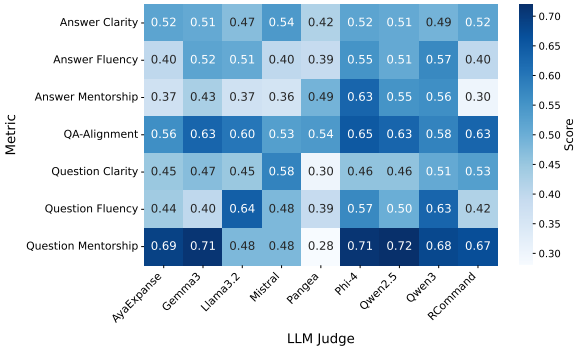

Figure 3: LLM-Human agreement scores. Qwen models and Phi-4 correlate the most with human results.

Reliability of Mentorship-Focused Evaluation. We assess the reliability of the proposed evaluation dimensions using human annotations. As expected, inter-annotator agreement is highest for linguistic metrics—*Fluency* (Q: 0.38, A: 0.42) and *Clarity* (Q: 0.43, A: 0.39)—which capture surface-level language properties. Importantly, task-oriented metrics—*QA Alignment* (0.39) and *Mentorship* (Q: 0.39, A: 0.35)—achieve stable, moderate agreement, reflecting the inherently interpretive nature of mentorship quality.

Agreement varies across models and languages. Among models, *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 Appendix D).

Overall, these patterns indicate that mentorship-focused evaluation is sensitive to both model behavior and linguistic context, mirroring real-world educational assessment and motivating future work on culturally aware and mentorship-specific evaluation frameworks.

Human–LLM Agreement in Automatic Evaluation. To assess the scalability of mentorship-focused evaluation, we compare LLM-based

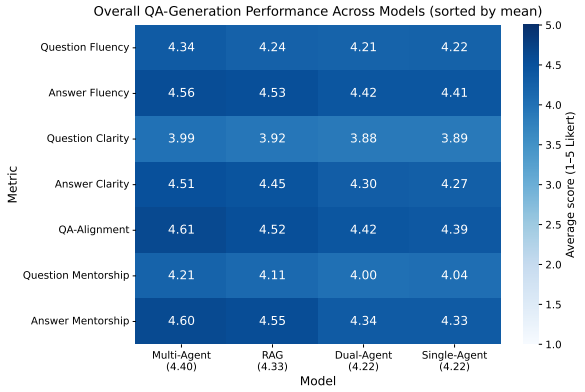

Figure 4: QA-generation performance across models and metrics (1–5 Likert). Models are sorted by mean.

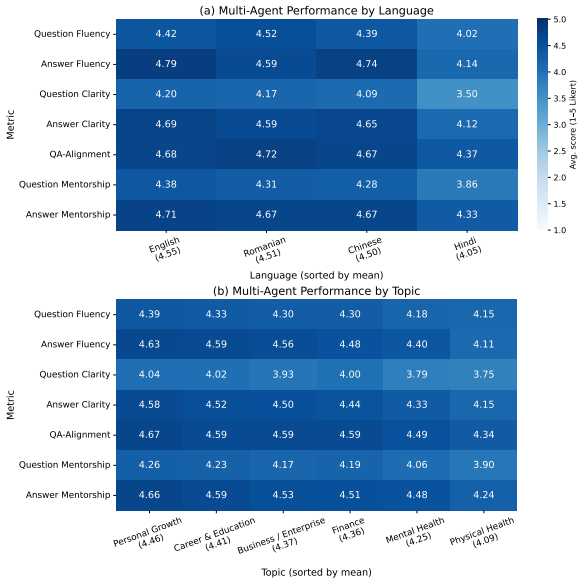

Figure 5: Multi-Agent performance across (a) language and (b) topic. Topics and languages are sorted by mean.

judges with human annotations (Figure 3). Agreement is highest for task-oriented metrics, with several judges (Qwen2.5, Qwen3, and Phi-4) reaching substantial agreement with human ratings (0.55–0.72), while linguistic metrics show moderate but stable alignment (0.40–0.57).

Qwen-family models (Qwen2.5, Qwen3) and Phi-4 consistently achieve the strongest agreement across metrics and languages, including lower-resource settings (results per language in Appendix D). At the same time, agreement varies substantially across judges, highlighting the challenge of automatically evaluating mentorship quality and motivating language- and metric-aware evaluation with multiple LLM judges rather than reliance on a single evaluator.

Comparison of QA-Generation Architectures. The Multi-Agent model achieves the strongest over-

all performance, ranking highest across all evaluation dimensions and obtaining the highest mean score (4.40), as shown in Figure 4. RAG follows closely (4.33), offering consistent improvements over the Single- and Dual-Agent baselines, particularly on *QA-Alignment*, *Question Mentorship*, and *Answer Mentorship*. In contrast, Single-Agent and Dual-Agent models perform nearly identically (4.22), suggesting that topic-aware chunking alone yields limited gains. Performance differences are most pronounced for *Mentorship* and *QA-Alignment* metrics, while *Fluency* remains relatively similar across models, highlighting the importance of agentic models for preserving mentorship quality. For detailed cross model and LLM-Judge results see Appendix D.2.

Qualitative comparisons of model outputs across different *Mentorship* scores are presented in Appendix D.2.

Analysis by Language and Topic. We analyze performance across languages and topics using the best-performing Multi-Agent model, with scores averaged over all LLM judges; cross model comparisons are reported in Appendix D.

Language Evaluation. As shown in Figure 5(a), *English achieves the highest overall performance (4.55)*, with *Romanian (4.51) and Chinese (4.50)* performing comparably across key metrics such as *QA-Alignment* and *Answer Mentorship*. In contrast, *Hindi scores lower overall (4.05)*, with the largest gaps in *Question Clarity* (3.50) and *Question Mentorship* (3.86), consistent with known tokenization and training limitations for Devanagari script. This aligns with MMLU results, where Hindi scores substantially lower than English (large-traversaal, 2025). Importantly, the Multi-Agent architecture substantially narrows this gap: compared to Single-Agent performance (QA-Alignment 3.96, Answer Fluency 3.72), *Multi-Agent improves Hindi scores to 4.37 and 4.14*, respectively. While disparities remain, these gains suggest that *agentic coordination can partially mitigate language-specific errors*.

Topic Evaluation. As shown in Figure 5(b), QA-generation performance varies substantially by topic. *Personal Growth achieves the strongest overall performance (4.46)*, leading across metrics such as *QA-Alignment* (4.67) and *Answer Mentorship* (4.66). *Career & Education (4.41) and Business/Enterprise (4.37)* follow closely, while *Physical Health exhibits the lowest overall scores (4.09)*, with notably lower *Question Clarity* (3.75) and *Question Mentorship* (3.90), reflecting the challenges posed by clinically dense language. In contrast, *Finance performs robustly*, achieving high *QA-Alignment* (4.47), likely due to its structured discourse. Overall, these results suggest that *agentic coordination is most beneficial for technically complex mentorship topics*, while offering limited gains for domains well represented in pre-training data.

Takeaways. Our results yield four key findings:
(1) Task-oriented metrics are reliable yet inherently more subjective than linguistic metrics.
(2) LLM-evaluation must be language- and metric-aware, motivating multi-judge strategies.
(3) Multi-Agent models consistently outperforms simpler architectures on mentorship and alignment dimensions.
(4) Agentic coordination is particularly beneficial for complex topics and lower-resource languages.

7 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—prioritizing guidance, clarity, and learning value beyond factual correctness—we find limitations in existing benchmarks, particularly in multilingual and long-context settings. Through a controlled comparison of Single-Agent, Dual-Agent, RAG, and Multi-Agent pipelines, we show that agentic coordination consistently improves mentorship quality, with the largest gains for complex topics and lower-resource languages, where it acts as a partial performance equalizer. Our analysis of automated evaluation further reveals substantial variation in alignment with human judgments across languages and metrics, highlighting the need for task-aware and language-sensitive evaluation. Together, these findings establish mentorship-focused QA as a rich and novel research direction and provide a multilingual benchmark for studying model architectures and evaluation methods in educational AI. We hope our work encourages the community to move beyond factual-driven evaluation toward models and metrics that better support learning, reflection, and guidance across cultures and languages.

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.

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.

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.

References

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. *arXiv preprint*.

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.

Team Cohere and Aakanksha. 2025. Command a: An enterprise-ready large language model.

Team Cohere & John Dang. 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.

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

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. 2025. Gemma-3-12b-it. https://huggingface.co/google/gemma-3-12b-it. Multilingual large vision-language model (Italian emphasis), hosted on Hugging Face.

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 8038–8046, Macau, China.

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.

Microsoft. 2025. Phi-4. https://huggingface.co/microsoft/phi-4. Large language model (Phi-4) hosted on Hugging Face.

Mistral AI. 2025. Ministral-8b-instruct-2410. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410. Language model, instruct-tuned, 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.

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.

S. Saadaoui, Y. Li, M. Zimmer, and L. Qin et al. 2025. Coordinated llm multi-agent systems for collaborative question answering. *Knowledge-Based Systems*, 302:109827.

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. *arXiv Preprint*.

Cameron R. Wolfe. 2024. Llm-as-a-judge. 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. 2025. Context selection and rewriting for video-based educational question generation. *arXiv Preprint*.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*.

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

B Video Sources

To see the full detailed source of all videos, see Table 3 and Table 4.

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[9]. 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[10], 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 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.

[9] https://www.youtube.com/@aclmentorship
[10] https://www.youtube.com/@MindArchitect

C 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.

C.1 Prompt Design

D Evaluation

Figure 10 shows the interface used in the human annotation process.

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

D.1.1 Inter-Annotator Agreement

Figure 11 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.

D.1.2 LLM-Human Agreement

LLM-Human Cohen kappa agreement scores per language are shown in Figure 12

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

All model performance across LLM-Judges is shown in Figure 13.

Qualitative results per model are shown below:

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, QAAlgn 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

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 6: Single-Agent prompt.

other students across various levels, including undergraduates, masters, and postdoctoral fellows.
Scores: QFlu 5, AFlu 5, QClar 4, AClar 4, QAAlgn 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, QAAlgn 5, QMen 5, AMen 3

Multi-Agent **Q:** What differences exist between PhD admission processes in various academic institutions?
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, QAAlgn 5, QMen 5, AMen 5

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 7: Dual-Agent prompt.

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

Language Evaluation. As shown in Figure 14, 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 architectural 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

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 8: RAG prompt.

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 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 15, 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 evaluation even with limited training examples.

As demonstrated in Figure 15, architectural complexity significantly impacts performance on challenging domains while showing minimal differen-

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 9: Multi-Agent prompt.

tiation 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.

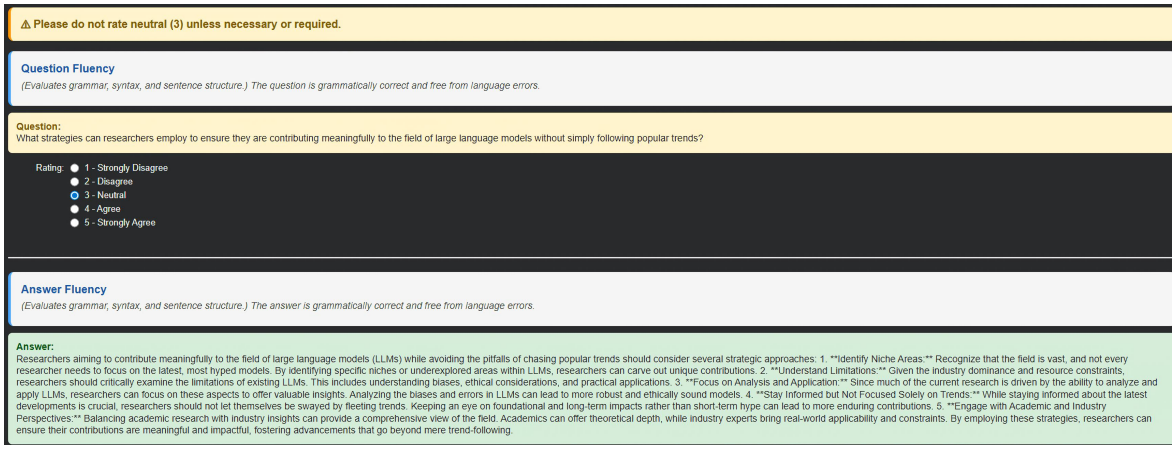

Figure 10: GUI of the Human Evaluation Process.

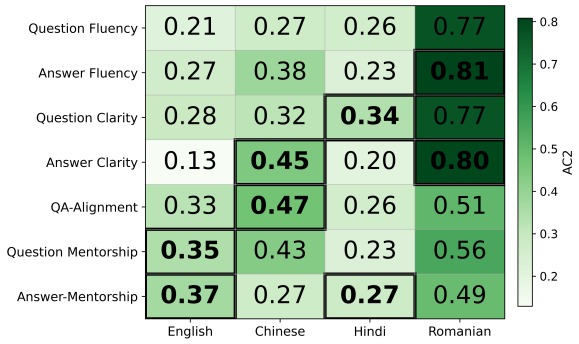

Figure 11: Inter-Annotator Agreement AC2 Scores by Language

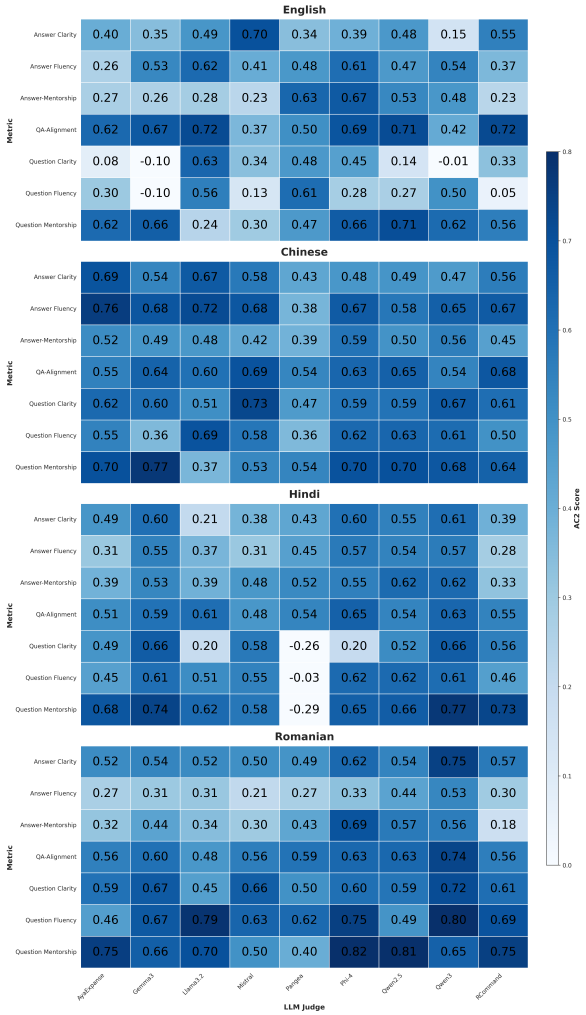

Figure 12: LLM-Human AC2 Agreement Scores

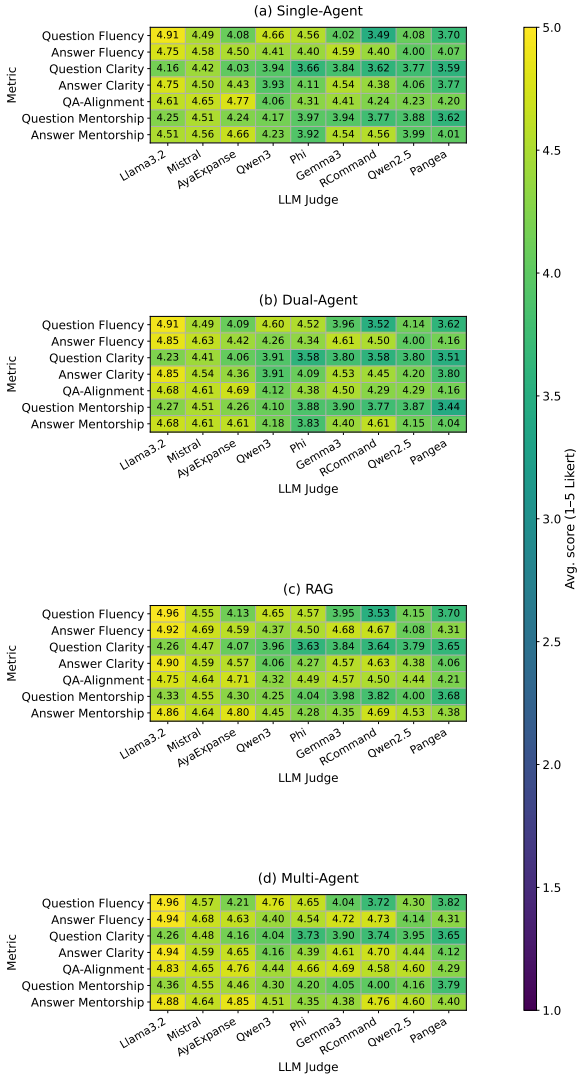

Figure 13: Judge-wise performance across QA-generation models: (a) Single-Agent, (b) Dual-Agent, (c) RAG, (d) Multi-Agent. Scores are averaged over metrics using a shared 1–5 Likert scale.

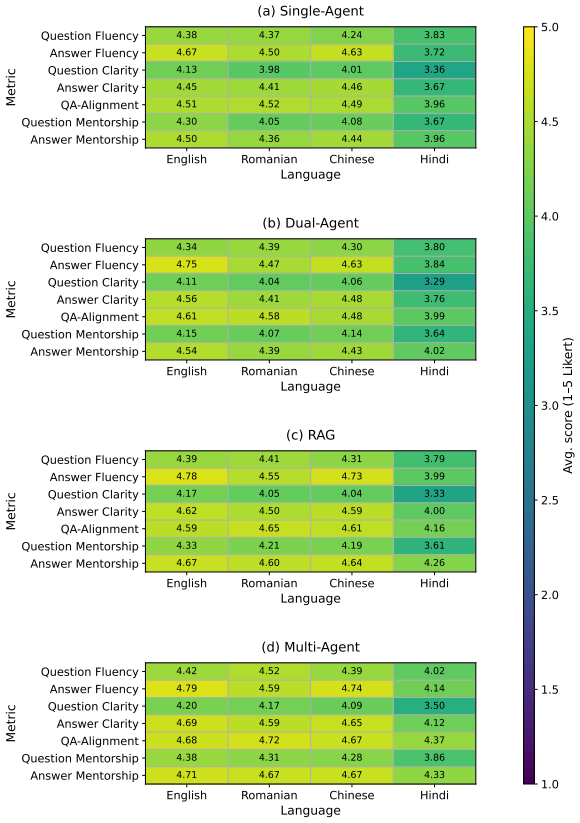

Figure 14: Language-wise performance across QA-generation models: (a) Single-Agent, (b) Dual-Agent, (c) RAG, (d) Multi-Agent. Scores are averaged over human and LLM judges (1–5 Likert).

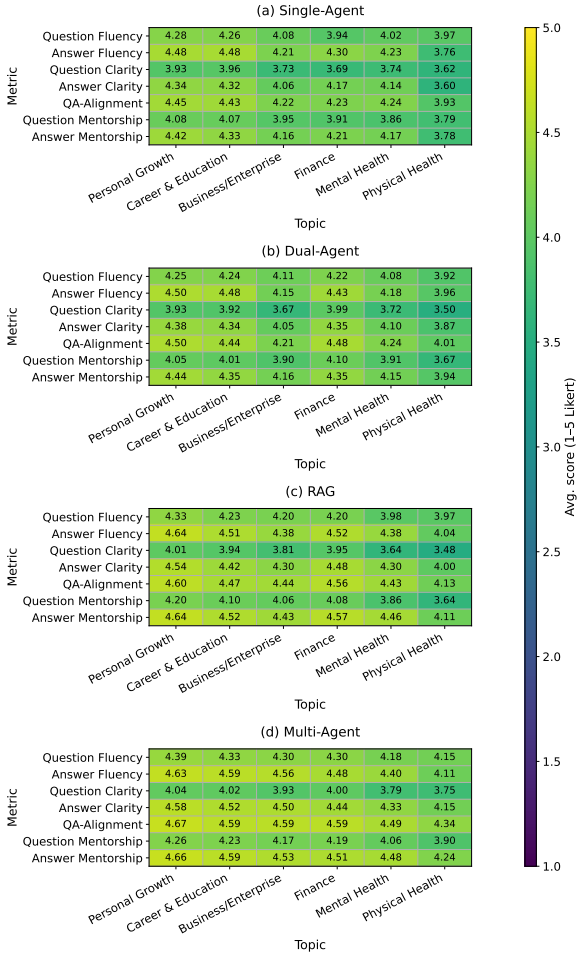

Figure 15: Topic-wise performance across QA-generation models: (a) Single-Agent, (b) Dual-Agent, (c) RAG, and (d) Multi-Agent. Scores are averaged over human and LLM judges (1–5 Likert scale).

Table 3: 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 4: Mentorship Videos (Continue)

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