MultiHint-9K: Orchestrating Human and AI Agency for Scalable Pedagogical Hint Generation

Kamyar Zeinalipour[1], Amir Sadeghi[1], Giovanni Angelini[2], Leonardo Rigutini[2], Marco Gori[1] and Marco Maggini[1]

[1]University of Siena, Siena, Italy
[2]Expert.ai, Modena, Italy

Abstract

Large Language Models offer transformative potential for personalized learning, yet they frequently fail at pedagogical scaffolding—leaking answers or using circular reasoning when generating hints. We present a framework that *orchestrates human and AI agency* to produce high-fidelity multilingual Socratic hints at scale. Human annotators define quality standards and calibrate automated critics; autonomous AI agents then enforce these standards through adversarial generation, critique, refinement, and recovery loops. This division of agency—humans set the rubric, AI operates at scale—produces **MultiHint-9K**: a dataset of 9,987 machine-verified, human-calibrated QA-Hint tuples across English, Italian, and Farsi, with a 97.5% yield rate. The pipeline's adversarial critics flagged 5,743 errors across generation attempts, including 3,131 circular reasoning and 495 answer leakage instances. We fine-tune five open-source models (7B–24B) via LoRA, achieving ∼67% ROUGE-L improvement over base models (all $p < 0.001$), with the largest gains in low-resource Farsi (∼97% ROUGE-L). A component contribution analysis demonstrates that removing any pipeline stage degrades yield from 98% to as low as 38%. We publicly release the dataset, models, and codebase at https://github.com/KamyarZeinalipour/MultiHint-9K.

Keywords

Multi-Agent Systems, Intelligent Tutoring, Hint Generation, Human-AI Agency, Pedagogical Scaffolding

1. Introduction

Effective pedagogical scaffolding—guiding learners toward understanding without revealing the answer—is foundational to Intelligent Tutoring Systems (ITS), where hints nudge students through their Zone of Proximal Development [1, 2, 3]. Traditional approaches rely on expert-authored content or data-driven methods that do not generalize across domains or languages [4, 5].

Large Language Models (LLMs) offer a compelling alternative, yet they frequently exhibit *answer leakage*—revealing the solution—and *circular reasoning*—restating the question—when generating hints [6, 7]. Prompt engineering can partially mitigate these failures in single-language settings [8], but provides no mechanism for systematic quality assurance at scale, and no prior work operates multilingually.

We propose a *division of agency* central to the HAI-Agency workshop theme: human annotators define quality rubrics and calibrate automated critics at the meta-level, while AI agents autonomously enforce these standards through adversarial generation-critique-refinement loops at scale. Concretely, we contribute: (1) a **human-calibrated multi-agent pipeline** catching 5,743 errors (495 leakage, 3,131 circular reasoning), with component analysis showing single-pass generation admits 33% erroneous hints; (2) **MultiHint-9K**, a multilingual dataset of 9,987 verified QA-Hint tuples with a cross-lingual error taxonomy; (3) **fine-tuned open-source models** (7B–24B, all $p < 0.001$); and (4) a **qualitative Refiner analysis** with real repair examples.

HAI-Agency: Workshop on Orchestrating Human and AI Agency for Proactive and Reflective Learning, co-located with the 27th International Conference on Artificial Intelligence in Education (AIED 2026), Seoul, Korea, June 27 – July 3, 2026.

Kamyar.zeinalipour@student.unisi.it (K. Zeinalipour); a.sadeghi@student.unisi.it (A. Sadeghi); postapergio@gmail.com (G. Angelini); lrigutini@expert.ai (L. Rigutini); marco.gori@unisi.it (M. Gori); marco.maggini@unisi.it (M. Maggini)

published 2026-08-27

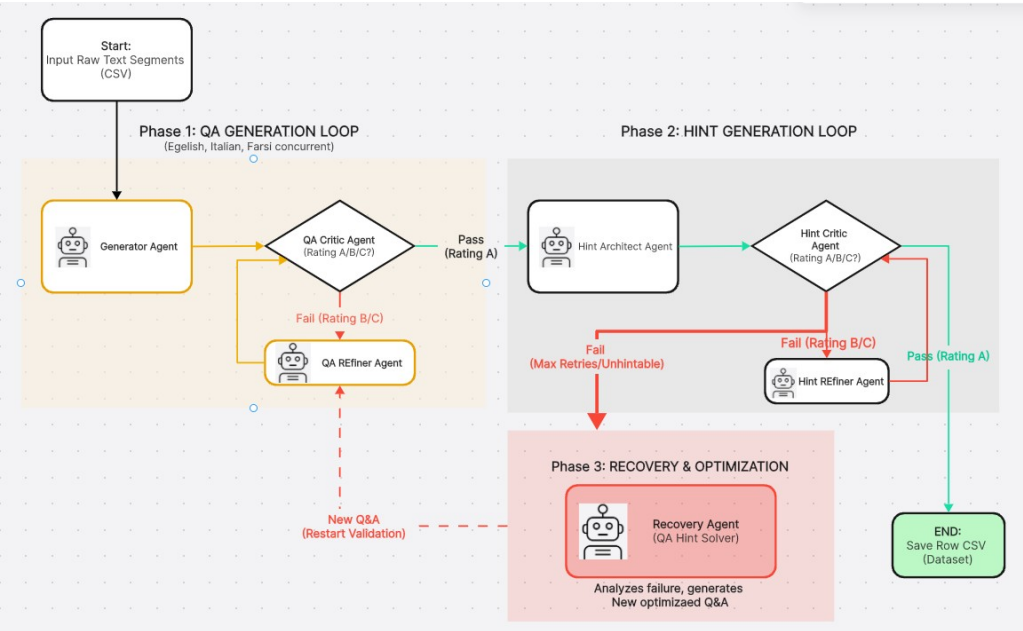

Figure 1: Pipeline architecture: Generator-Critic-Refiner loop with Recovery Agent, and a real generation example.

2. Related Work

Hint generation and LLM scaffolding. Hint generation in ITS has evolved from expert-authored sequences [3] to data-driven approaches mining student interactions [5]. Kochmar et al. [4] advanced automated pedagogical intervention generation but do not generalize across languages. The TriviaHG dataset [9] demonstrates factoid hint feasibility but relies on single-model prompting without adversarial quality assurance. Jangra et al. [6] survey hint generation—including Socratic variants, where hints must scaffold without revealing—identifying answer leakage and circular reasoning as persistent open challenges. Pardos and Bhandari [8] showed that ChatGPT can produce algebra hints with learning gains in single-language settings; Mozafari et al. [7] identified significant quality variation across LLMs; Sun et al. [10] proposed AutoHint for automatic prompt optimization. Our prior work spans related educational content formats: clue generation [11], multilingual QA [12, 13], Italian [15] and Arabic educational content [14]; none provide adversarial quality assurance for Socratic hints.

Multi-agent systems and human-AI agency. Sengupta et al. [16] demonstrated MAG-V, showing adversarial agent interactions improve synthetic data quality; Li et al. [17] surveyed retrieval-augmented generation for educational applications. Our pipeline extends the Generator-Critic-Refiner paradigm with a Recovery Agent grounded in human calibration. The orchestration of human and AI agency is increasingly recognized as central to effective educational technology [18]: Wood et al.'s [1] scaffolding metaphor directly motivates our hint design, while Zimmerman's [19] framework supports a meta-level human role that calibrates rather than annotates at scale.

3. The Proposed Framework

3.1. Pipeline Architecture

Our system is a three-phase pipeline across three parallel language streams (English, Italian, Farsi), totaling 21 concurrent agents (7 per language). Each stream contains a Generator, Critic, and Refiner for both QA and Hint phases, plus a shared Recovery Agent. Figure 1 illustrates the architecture.

3.2. Phase 1: QA Synthesis

A *Generator* agent scans Wikipedia articles—selected for broad cross-lingual coverage, free availability, and established use in multilingual QA benchmarks; filtered to the top tertile by length within Science, History, and Geography—to synthesize complex QA pairs targeting causally linked entities. A *Critic* agent audits each candidate via a structured Chain-of-Thought prompt for: (a) hallucination; (b) ambiguity; (c) truncation; and (d) format violations. Failed candidates go to a *Refiner* for repair (up to $N = 3$ iterations; empirically justified: Loop 2 contributes 10.1% additional QA samples, Loop 3 only 2.4%). Open-source models (7B–24B) were selected over larger closed alternatives to support local deployment and eliminate API dependency; Table 1 shows no single model dominates across roles and languages, motivating the multi-agent design over monolithic approaches.

3.3. Phase 2: Hint Scaffolding

Valid QA pairs enter the hint scaffolding protocol. A *Hint Generator* creates Socratic hints via **orthogonal attribute extraction**—describing the answer using semantic features (temporal, geographic, causal) absent from the question. A *Hint Critic* evaluates each candidate via Chain-of-Thought, rejecting for: **circular reasoning** (restating the question), **answer leakage** (revealing the solution), **hallucination** (unsupported facts), or **vagueness** (insufficient scaffolding), and passes the specific error code to the *Hint Refiner* for targeted repair ($N = 3$ iterations).

3.4. Phase 3: Recovery

When the Hint Refiner fails after maximum iterations, a **Recovery Agent** re-scans the source text and generates a new QA-Hint tuple targeting an alternative entity (up to $K = 2$ pivots), converting potential waste into verified training data. The Recovery Agent was activated 1,903 times—1,397 for Farsi alone—evidencing a significant "alignment tax" for non-Latin languages.

3.5. Agent Quality Metrics

We use two standard information-retrieval metrics computed on human-rated samples for agent selection. **A-Recall** (recall over Class-A human-rated samples) measures the proportion of human-confirmed acceptable samples also accepted by the LLM Critic; high A-Recall ensures good content is not over-rejected. **A-Precision** (standard precision) measures the proportion of Critic-accepted samples confirmed by humans. We prioritize A-Recall for Critic selection because false negatives are recoverable through the Refiner, while false positives propagate errors into the dataset. Selected Critics achieve A-Precision of 71%–97% across tasks and languages, with the adversarial pipeline structure providing a structural check against correlated failures.

4. Human-Calibrated Agent Selection

To assign optimal agents, we conducted a two-stage evaluation using 100 source texts per language. Two annotators per language (native speakers for Italian and Farsi; IELTS Band 7+ for English) independently rated outputs from five candidate models—DeepSeek-Chat (DS-C), DeepSeek-Reasoner (DS-R), Mistral-Large, Gemini-1.5-Flash, and GPT-4o—on a 3-point Likert scale: 3 (Acceptable), 2 (Partially Acceptable), 1 (Unacceptable). Initial agreement was substantial (Cohen's κ [20]: En= 0.74, It= 0.73, Fa= 0.69, per Landis and Koch [21]); disagreements were resolved by consensus. This evaluation provides an **independent quality gate**—breaking circularity between LLM-generated content and LLM-based critique—and exemplifies the HAI orchestration design: humans calibrate quality standards once at the meta-level, AI agents enforce them at scale.

Stage 1 (Generator Selection): The highest-scoring model was selected per task and language: **Gemini** for QA in English (2.96) and Farsi (2.90), **Mistral** for Italian QA (2.77), **GPT-4o** for English

Table 1
Human-Calibrated Agent Selection. *Left*: Generator quality (mean human score, 1–3). *Right*: Critic reliability (A-Recall %). **Bold** = selected agent.

Model	Stage 1: Generator (Human Score)						Stage 2: Critic (A-Recall %)					
	QA Gen			Hint Gen			QA Critic			Hint Critic		
	En	It	Fa	En	It	Fa	En	It	Fa	En	It	Fa
DS-Chat	2.83	2.73	2.88	2.48	**2.62**	**2.48**	88.5	**95.4**	90.2	85.7	77.9	32.3
DS-Reasoner	2.86	2.69	2.79	2.70	2.50	2.42	83.3	89.7	92.4	88.6	91.2	71.0
GPT-4o	2.81	2.64	2.69	**2.71**	2.48	2.35	83.3	85.6	92.9	91.4	**97.0**	82.3
Mistral	2.72	**2.77**	2.65	2.59	2.41	2.33	**94.3**	93.1	93.5	**97.1**	95.6	**95.2**
Gemini	**2.96**	2.73	**2.90**	2.46	2.09	2.34	89.1	91.4	**94.0**	97.1	91.2	71.0

Table 2
MultiHint-9K dataset yield by language.

Language	Articles	Tuples	Yield
English	3,421	3,327	97.2%
Italian	3,800	3,793	99.8%
Farsi	3,017	2,867	95.0%
Total	**10,238**	**9,987**	**97.5%**

hints (2.71), and **DS-Chat** for Italian (2.62) and Farsi (2.48) hints. No single model dominates across all languages and tasks, empirically justifying specialized, language-specific agents.

Stage 2 (Critic Selection): Using the same human-rated subset as ground truth, we measured each model's A-Recall. The highest A-Recall model was selected as Critic for each task and language, and also assigned to the Refiner and Recovery Agent roles (same generative task). Table 1 reports all results.

5. Dataset Characteristics

5.1. Source Corpus and Data Yield

We sourced **10,238 Wikipedia articles** (English: 3,421; Italian: 3,800; Farsi: 3,017). Table 2 shows the pipeline achieved a 97.5% overall yield, producing 9,987 verified QA-Hint tuples (**MultiHint-9K**). The lower Farsi yield (95.0%) reflects the increased difficulty models face with non-Latin semantic structures, consistent with alignment challenges observed in Farsi educational content generation [13]. While question length varies dramatically across languages (Italian $\mu = 40.0$ vs. English $\mu = 12.1$ tokens), hint length remains stable ($\mu \approx 23$–28 tokens), suggesting that effective Socratic scaffolding requires consistent information density regardless of language.

5.2. Adversarial Filtering and Error Taxonomy

The system executed **31,600 agent activations** to produce 9,987 verified tuples—a Quality Verification Ratio of 3.16:1. Farsi consumed the highest resources (12,237 activations) despite producing the fewest tuples, confirming the "alignment tax" for non-Latin languages. Table 3 details the errors caught by Critic agents. In hint generation, **circular reasoning** was the dominant failure (3,131 rejections), with Farsi models particularly prone (1,815 instances). Italian showed high rates of both leakage (322) and vagueness (282), suggesting difficulty balancing hint specificity. Question type distributions reflect typological differences: English favors "What" (58%), Farsi is entity-focused ("Who" 33%, "Which" 29%), and Italian is dominated by "Quale" (57%) and "Come" (29%)—patterns that emerged naturally from generator-Wikipedia interaction, evidencing cross-lingual adaptability rather than a fixed question template.

Table 3
Error taxonomy: errors detected by Critic agents during QA and Hint generation.

QA Errors	En	Fa	It	Tot.	Hint Errors	En	Fa	It	Tot.
Truncation	237	147	144	**528**	Circular Reas.	850	**1815**	466	**3131**
Inexact Ref.	269	35	108	412	Ans. Leakage	103	70	322	495
Format	191	1	24	216	Vagueness	96	16	282	394
Extra Content	112	5	70	187	Bad Reference	12	48	104	164
Mismatch	7	23	73	103	Hallucination	49	56	8	113

Table 4
Base (B) vs. fine-tuned (FT) ROUGE-L per language and mean BERTScore across languages. All improvements $p < 0.001$ (Bonferroni-corrected Wilcoxon).

	En ROUGE-L		Fa ROUGE-L		It ROUGE-L		Avg BRT	
Model	B	FT	B	FT	B	FT	B	FT
Llama-8B	.21	**.32**	.15	**.31**	.25	**.33**	.72	**.77**
Qwen-14B	.26	**.29**	.21	**.33**	.23	**.31**	.73	**.76**
Mistral-7B	.20	**.30**	.16	**.32**	.20	**.34**	.71	**.77**
Nemo-12B	.16	**.31**	.12	**.30**	.18	**.35**	.69	**.77**
Mistral-24B	.20	**.31**	.16	**.33**	.17	**.32**	.70	**.77**
Average	.21	**.31**	.16	**.32**	.21	**.33**	.71	**.77**

6. Fine-Tuning Experiments

6.1. Experimental Setup

MultiHint-9K was split into training (9,687 samples) and a language-stratified test set (300 samples; 100 per language). We fine-tuned five open-source models—**Llama-3.1-8B**, **Qwen2.5-14B**, **Mistral-7B-v0.3**, **Mistral-Nemo-12B**, **Mistral-Small-24B**—using LoRA [22] ($r = 32$) on a 4×A6000 node for 3 epochs (lr $= 10^{-4}$, batch 32), evaluated with BERTScore [23] and ROUGE-L [24].

6.2. Results

Table 4 presents the results. Fine-tuning produced consistent ROUGE-L and BERTScore improvements across all models and languages (Wilcoxon signed-rank, $p < 0.001$, 300 paired samples; all significant after Bonferroni correction at $\alpha = 0.003$). Global average ROUGE-L improved from 0.19 (base) to 0.32 (fine-tuned), a ∼67% gain.

Mistral-Nemo-12B emerged as the efficiency champion: despite the weakest baseline (ROUGE-L 0.12–0.18), it achieved the largest post-fine-tuning gains and reached parity with Mistral-Small-24B, demonstrating that alignment with verified pedagogical data matters more than model scale. **Farsi** recorded the largest ROUGE-L improvements (∼97%), substantially exceeding English (∼52%) and Italian (∼60%). This asymmetry directly reflects the alignment tax documented in Section 5: Farsi required 39% recovery activations vs. 4.5% for English—pre-training data scarcity creates larger headroom for improvement through high-quality adversarially-filtered synthetic data. BERTScore gains are modest (∼6–8%) due to known metric saturation at high values, but confirm semantic alignment is preserved. Without adversarial critique, ∼33% of hints would contain errors (Table 5), establishing the single-pass baseline. ROUGE-L and BERTScore remain *proxy* metrics; a controlled student study measuring learning gains is essential future work.

6.3. Component Contribution Analysis

Table 5 analyzes component contributions using recorded loop indices (*projected estimates* from pipeline logs, providing lower-bound estimates of each stage's value).

Table 5
Component contribution: projected yield under different pipeline configurations.

Configuration	Est. Yield	Error Risk
Generator only (no Critic)	98.0%	33.1% of hints contain errors
Generator + Critic (no Refiner)	37.7%	Low errors, massive waste
Gen + Critic + Refiner (no Recovery)	82.2%	Low errors, moderate waste
Full Pipeline	**98.0%**	**Only 2.5% discarded**

Table 6
Critic rejection and Refiner repair examples (English, MultiHint-9K).

Role	Content
Ex. 1 — Q: "What year marked the end of the Victorian era?" A: 1901	
Hint v1	"Consider the year associated with the death of Queen Victoria, marking a significant transition between eras."
Critic	**REJECTED** — ERR_CIRC: Restates "Victorian era" without new information.
Hint v2	"Think about when Edward VII was influenced by continental European art, coinciding with a significant historical change."
Critic	**ACCEPTED** — Orthogonal attributes (Edward VII, continental art).
Ex. 2 — Q: "Who is the wife of the Hindu god Shiva?" A: Parvati	
Hint v1	"Recall the goddess described as the reincarnation of Sati, who is also the wife of the Hindu god Shiva."
Critic	**REJECTED** — ERR_CIRC: Restates the question ("wife of Shiva").
Hint v2	"Remember that she is the daughter of the mountain-king Himavan and queen Mena."
Critic	**ACCEPTED** — Genealogical scaffolding without naming the answer.

Without the Critic, 33.0% of hints pass with errors undetected; without the Refiner, yield drops to 37.7%; the Recovery Agent salvages 1,614 samples (16.1%). Diminishing returns justify $N = 3$: Loop 2 contributes 10.1% (QA) and 22.8% (hints), while Loop 3 adds only 2.4% and 10.2% respectively.

6.4. Qualitative Examples

Table 6 illustrates the Critic's error detection and Refiner's repair behavior on two English examples.

7. Discussion

Human-AI agency orchestration. Our framework exemplifies a productive division of agency: **human agency** at the meta-level (defining rubrics, calibrating critics via A-Recall) and **AI agency** at the execution level (31,600 activations for 10K tuples). A key finding—no single LLM dominates across all languages and roles (Table 1)—empirically justifies the multi-agent architecture and demonstrates that effective HAI orchestration goes beyond HITL annotation: humans set the quality standard once; AI enforces it at scale.

Limitations. (1) Component contribution uses logged indices, not re-runs (lower-bound estimates). (2) Proxy metrics are used; a controlled student study is essential. (3) Human calibration covers ∼3% of outputs. (4) No direct comparison to prior systems [4, 9] due to domain differences. (5) MultiHint-9K covers single-turn factoid Wikipedia hints; closed-model APIs are required for some roles, mitigated by our public code/data release. Future work: controlled student studies, multi-turn scaffolding, and expansion to additional languages and STEM domains.

Ethical considerations. Wikipedia biases may propagate into hints; auditing is needed before classroom deployment. Cross-linguistic disparities (Farsi: 39% recovery vs. English: 4.5%) reveal unequal

LLM pedagogical support across languages, raising fairness concerns. MultiHint-9K produces *candidate* content requiring teacher review before student exposure.

8. Conclusion

We presented a framework orchestrating human and AI agency for scalable multilingual hint generation. The pipeline caught **5,743 errors** while maintaining a 97.5% yield (vs. 33% single-pass error rate); each stage is necessary, with yield dropping to 37.7% without the Refiner. Fine-tuned open-source models (7B–24B) achieved significant improvements across all three languages ($p < 0.001$), with Nemo-12B gaining $\sim$110% in ROUGE-L. MultiHint-9K, models, and codebase are publicly released.

Declaration on Generative AI

GPT-4o and Gemini assisted with writing and proofreading of this paper. GPT-4o, Gemini, DeepSeek, and Mistral were also used as pipeline components for MultiHint-9K data generation. All content was reviewed and validated by the authors, who take full responsibility.

References

[1] D. Wood, J. S. Bruner, and G. Ross. The role of tutoring in problem solving. *J. Child Psych. & Psychiatry*, 17(2):89–100, 1976.

[2] L. S. Vygotsky. *Mind in Society*. Harvard University Press, 1978.

[3] K. VanLehn. The relative effectiveness of human tutoring, ITS, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[4] E. Kochmar et al. Automated data-driven generation of personalized pedagogical interventions in ITS. *Int. J. AI in Education*, 32:323–349, 2022.

[5] T. Barnes and J. Stamper. Toward automatic hint generation for logic proof tutoring. In *Proc. ITS*, pages 373–382, 2008.

[6] A. Jangra, J. Mozafari, A. Jatowt, and S. Muresan. Navigating the landscape of hint generation research. *Trans. ACL*, 13:505–528, 2025.

[7] J. Mozafari, A. Abdallah, B. Piryani, and A. Jatowt. Exploring hint generation approaches in open-domain QA. *arXiv:2409.16096*, 2024.

[8] Z. A. Pardos and S. Bhandari. ChatGPT-generated help produces learning gains equivalent to human tutors on math. *PLoS ONE*, 19(5):e0304013, 2024.

[9] J. Mozafari, A. Jangra, and A. Jatowt. TriviaHG: A dataset for automatic hint generation from factoid questions. In *Proc. SIGIR*, pages 2060–2070, 2024.

[10] H. Sun et al. AutoHint: Automatic prompt optimization with hint generation. *arXiv:2307.07415*, 2023.

[11] A. Zugarini, K. Zeinalipour, S. S. Kadali, M. Maggini, M. Gori, and L. Rigutini. Clue-Instruct: Text-based clue generation for educational crossword puzzles. In *Proc. LREC-COLING 2024*, pages 3853–3862, 2024.

[12] K. Zeinalipour, Y. G. Keptiğ, M. Maggini, and M. Gori. Automating Turkish educational quiz generation using LLMs. In *ISPR*, pages 246–260, 2024.

[13] K. Zeinalipour et al. PersianMCQ-Instruct: Generating multiple-choice questions in Persian. In *Proc. Workshop on LMs for Low-Resource Languages*, 2025.

[14] K. Zeinalipour et al. LLM-driven Arabic educational crossword generation. In *Proc. WLLRL*, pages 479–495, 2025.

[15] K. Zeinalipour et al. Italian crossword generator for education. *arXiv:2311.15723*, 2023.

[16] S. Sengupta et al. MAG-V: A multi-agent framework for synthetic data generation and verification. *arXiv:2412.04494*, 2025.

[17] Z. Li et al. Retrieval-augmented generation for education: A systematic survey. *Computers and Education: AI*, 8:100417, 2025.

[18] M. N. Giannakos et al. The promise and challenges of generative AI in education. *Behaviour & Information Technology*, 44(11):2518–2544, 2024.

[19] B. J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory into Practice*, 41(2):64–70, 2002.

[20] J. Cohen. A coefficient of agreement for nominal scales. *Educ. Psych. Measurement*, 20(1):37–46, 1960.

[21] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[22] E. J. Hu et al. LoRA: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022.

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT. In *Proc. ICLR*, 2020.

[24] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL Workshop: Text Summarization*, pages 74–81, 2004.