Edustories: A Collection of Real-world Case Studies from Classroom Practices

Anonymous ACL submission

Abstract

Despite the widely recognized potential of AI applications in education, most prior work has focused on individualized student assistance. In contrast, the majority of educational practice worldwide still takes place in collective classroom settings. To enable AI research grounded in classroom practice, we present Edustories, a dataset of 1,492 teacher-written case studies describing real elementary and high-school classroom situations involving challenging student behavior, pedagogical interventions, and their outcomes.[1] Among its applications, Edustories enables a systematic evaluation of large language models in their ability to provide practising teachers with feedback on their interventions. By comparing the latest models from four language-model families with standardized expert assessments, we demonstrate that current models fall short of human expertise in predicting classroom outcomes; The strongest models reach 58% accuracy compared to 64% of human experts, underscoring the limitations but also the emerging potential of current AI as teaching assistants rather than replacements.

1 Introduction

Artificial intelligence has a large potential to support and improve educational practice with over one third of teachers already using AI in their daily practice to prepare materials or personalize exercises (Walton Family Foundation, 2025). Prior work has demonstrated that AI-driven systems can accelerate learning and improve outcomes in a variety of educational domains, including intelligent tutoring systems, automated feedback for programming and mathematics, and adaptive learning platforms (Matos et al., 2025). However, much of this progress has focused on personalized tutoring, where models interact with a single learner and generate personalized explanations, hints, or corrections (Wang et al., 2024a).

In contrast, collective classroom instruction—in which a teacher simultaneously engages with many students—remains the dominant educational paradigm in elementary and secondary schools. Practicing teachers in such settings face challenges that go beyond individual learning trajectories, including classroom management, social dynamics among students, and the need to balance instructional goals with behavioral interventions (Kearney, 2019). Despite its central role in everyday education, relatively little prior work has aimed to develop automated systems that explicitly support teachers in managing and reflecting on collective lessons (Wang et al., 2024a).

We argue that this gap can be attributed to two main factors. First, collective lessons are inherently more complex than one-on-one educational interactions, as they involve multiple actors, social relationships, and context-dependent dynamics that are difficult to model computationally (Toom and Husu, 2021). Second, the collection of data documenting real classroom situations poses significant challenges: strict ethical and legal requirements for protecting the anonymity of students and practicing teachers limit the availability of detailed, real-world descriptions without compromising authenticity or completeness (Varlamis, 2025). These constraints likely explain why, to best of our knowledge, there exists no other publicly available textual dataset documenting situations arising in collective classroom instruction.

To address this gap and to support the development of AI technologies for assisting teachers in collective lessons, we introduce Edustories, a dataset of 1,492 teacher-authored descriptions of challenging situations encountered during classroom instruction. Each case study includes a narrative description of the situation, the teacher's chosen response or intervention, and the implied out-

[1] The Edustories dataset is available at https://huggingface.co/datasets/authoranonymous321/Edustories

come, reflecting the teacher's own assessment to what extent the intervention was successful.

Edustories opens up new research towards evaluating and improving AI systems designed to assist teachers in daily practice. In this paper, we focus on one particularly impactful application: providing feedback to teachers' proposed solutions of classroom situations by modeling their most likely outcome. Using the Edustories dataset for evaluation, we find that modeling qualitative outcomes of classroom interventions remains challenging even for strongest LLMs, particularly in distinguishing between short-term and long-term success, with 5 out of the 6 evaluated models underperforming the lower-bound of standardised expert assessment.

1.1 Related collections

Classroom behaviour Several datasets study classroom behaviour using video recordings of teaching practice, including MM-TBA (Huang et al., 2025), TBU (Cai et al., 2025), and FSCB (Lin et al., 2024), with the latter two also providing explicit behaviour labels. These datasets capture what happens in the classroom at an observable level, but they are limited to video data. As a result, they do not represent deeper situational factors captured in Edustories – such as students' prior history, earlier interactions, or how teachers perceive and interpret classroom events. Such information is essential for pedagogical decision-making and providing a meaningful feedback.

Teacher reflections Teacher reflection has been explored through textual datasets aimed at analysing affective and cognitive aspects of teaching practice. The CEReD dataset (Štefánik and Nehyba, 2021) consists of teacher reflection texts annotated for emotional and reflective elements and has been used for emotion classification and reflective analysis (Nehyba and Štefánik, 2022; Zhang et al., 2024), often with the goal of providing feedback to teachers on their practice (Hofmann et al., 2025; Zhang, 2024). However, existing datasets are designed for classification and do not preserve complete classroom narratives. In contrast, Edustories provides holistic case descriptions that retain the rich semantics and narrative of class situations.

Narrative datasets Narrative datasets that model the evolution of situations and the consequences of actors' actions have been developed in several non-educational domains. Examples include datasets focused on moral decision-making and plausibility reasoning (Emelin et al., 2021; Bae et al., 2025), historical narratives (Leteno et al., 2025), social interactions and norms (Forbes et al., 2020), and ethical judgments (Hendrycks et al., 2021a). While these resources share a narrative structure similar to Edustories, they are not directly relevant for educational practice.

2 The Edustories Dataset

Origin and Preparation The case studies contained in the Edustories dataset were written by trainee teachers during their teaching practice at Czech elementary and high schools, coordinated between 2023 and 2026. These trainee teachers are finishing university students completing classroom practice as the last part of their curricula, collecting their case studies with the primary goal of receiving feedback in their collective reflection sessions that follow their teaching practice.

Each case study is recorded in a structured format designed to capture key aspects of classroom situations. Mainly, the collected stories include (1) a narrative description of the situation, (2) an anamnesis providing background information about the student playing a central role in the story, (3) the problem the teacher had to address, (4) the intervention applied by the teacher as a response to the situation, and (5) a description of the perceived outcome of the intervention.

The collected stories underwent a peer-reviewed process where the dataset was further enriched with categorical features including outcome categories reflecting the authors' self-assessment of intervention success: *failure*, *Partial or short-term success*, and *Long-term success*. Other categorical variables provided directly by each story's author include the age of the student, their hobbies, and the presence of any reported disorders or diagnoses.

To make the dataset accessible to the broader AI and NLP research community, the free-text components of Edustories (description, anamnesis, problems, and solutions) were translated from Czech to English. Translation was performed using DeepL, selected as the best-performing translation system from a set of commercial translation services based on an evaluation of 50 expert-verified translations.

Consent and Anonymization To protect the privacy of all participants, the free-text case descriptions underwent a two-stage anonymization process. First, all personal names and explicit (non-relative) references to dates, times, and locations were automatically identified and randomized us-

Feature	Num. filled	Avg. words	Avg. sentences	Num. categories	Most common
Description	1492	141.39	9.41	–	–
Anamnesis	1491	79.55	5.86	–	–
Intervention	1492	134.99	8.21	–	–
category	1475	–	–	15	interview (30.35%)
Outcome	1492	82.78	5.26	–	–
category	1475	–	–	3	long-term success (44.63%)
Student age	1455	–	–	–	12 years
Hobbies	1365	4.75	–	–	computer games (8.17%)
Diagnosis	583	–	–	72	ADHD (22.52%)
Disorders	908	–	–	103	lying (24.66%)

Table 1: Overview of statistics of the main features of Edustories data collection (1,492 samples in total).

ing the UFAL Named Entity Recognition service tailored for Czech (Ševčíková et al., 2007). Second, the anonymized texts were manually reviewed during peer-reviewed quality control and augmentation to ensure that no identifying information remains. All authors of Edustories case studies voluntarily provided written informed consent[2] permitting the further use of their data in anonymized form.[3]

3 Experiments: Predicting Invervention Success

One of the most direct and impactful applications of AI for improving collective classroom instruction is the provision of feedback to teachers on their instructional and intervention strategies. In the absence of dedicated benchmarks, teachers seeking to benefit from AI-generated feedback must rely on a trial-and-error process, iteratively incorporating suggestions into subsequent classroom practice. This approach has notable limitations: if the quality of the feedback is uncertain, its application may temporarily degrade instructional effectiveness and adversely affect students' learning experiences.

The Edustories dataset provides explicit annotations of intervention outcomes, enabling a systematic evaluation of the quality of AI-generated feedback on teachers' proposed intervention strategies. In this study, we use these outcome annotations to assess the performance of six state-of-the-art language models from five model families, and we compare their predictions to standardized evaluations provided by educational experts.

3.1 Experimental Setup

Data For each case study, we construct evaluation prompts by concatenating the *Description*, *Anamnesis*, and *Intervention* fields. We experiment with five alternative prompt formulations and select the prompt that achieves the highest overall accuracy across models as the final evaluation prompt. We note that differences between prompt variants fall within the evaluation confidence interval (±1.6%), indicating minor sensitivity to prompt choice. Thus, to favour transparency and reproducibility, we use a matching prompt for all models evaluations (detailed in Appendix A.2).

Models We evaluate a set of open-source language models that represent the state of the art across a range of sizes and model families. As a proxy for model quality, we select models that lead their respective size categories on the MMLU-Pro benchmark (Hendrycks et al., 2021b; Wang et al., 2024b). The evaluated models span the most recent models from Llama 3, Qwen 3, Mistral v0.3 and Olmo 3 families; concrete references can be found in Appendix A.2.1. As we observe no consistent relationship between model size and performance, our evaluations include both larger and smaller models.

Expert Annotations As a human reference, we compare model performance against evaluations provided by five educational experts. Our experts are experienced educational practitioners familiar with the Behavior Management Frameworks detailed below. The experts assessed 310 selected case studies from Edustories, evaluating the proposed *Intervention* given the corresponding *Description* and *Anamnesis*. Each intervention was evaluated along three dimensions: (1) **Humanistic orientation** (scale -2 to +2), reflecting emphasis on empathy, dialogue, and student needs versus reinforced control; (2) **Reactivity** (scale -2 to +2), capturing *reactive* (cf. *preventive*) strategies; and (3) **Systematicity** (scale -2 to +2), distinguishing generalizable strategies with contextual adaptation from incident-specific actions.

These dimensions are grounded in Behavior Management Frameworks, namely Social and Emo-

[2] In the name of the authors of Edustories case studies, we reserve the right to withdraw specific samples from Edustories dataset in every case when the original author wishes to do so.

[3] The proposal of data collection first underwent review by the ethical committee of the coordinating organisation (to be specified in a non-anonymous version) with a positive result.

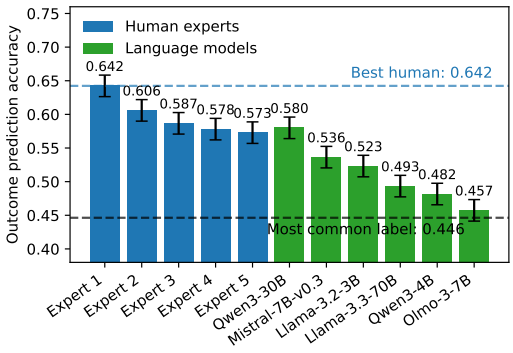

Figure 1: Accuracy of outcome prediction for interventions and case studies of Edustories dataset, for (blue) appropriateness evaluation of human experts and (green) direct prediction of selected language models. Evaluation on n=310 samples to allow comparability with judgements of human experts (§3.1). Confidence interval estimates for $\alpha = 0.1$ covering ±1.6% ranges.

tional Learning (Cipriano et al., 2023), Nonviolent Communication (Batūraitē-Bunka et al., 2024), Positive Behavioral Interventions and Supports (Bradshaw et al., 2012), and Restorative Practices (Fronius et al., 2019). Based on these aspects, experts finally rated the overall **Appropriateness** of each intervention on a scale from -2 to +2.

To align expert judgments with the Edustories outcome labels, we map Appropriateness scores to outcome categories as follows: a score of -2 correspond to *failure*, a scores of -1 and 0 corresponds to *partial or short-term success*, and scores of +1 and +2 correspond to *long-term success*.

3.2 Results

Figure 1 compares outcome prediction accuracy between human experts and selected language models. Among all evaluated models, only one model (Qwen-3-30B) reaches performance overlapping with human expert accuracy, achieving 0.580 compared to 0.573–0.587 for the lower-performing experts. Qwen-3-30B also exhibits a clear margin over all other language models.

Models' performance does not appear to depend on model size, which has positive implications for practical deployment; For example, the larger Llama-3.3-70B underperforms the smaller Llama-3.2-3B, suggesting that architectural choices, training data, alignment, and other training refinements play a more important role than scale alone. Beyond size, outcome prediction ability may relate to general reasoning capability, as the two best-performing models also show particularly strong performance on reasoning-focused tasks.

Comparative analysis A closer comparison of the best-performing model and human expert (Appendix A.1) reveals qualitatively different error patterns. Human experts' errors predominantly arise from assigning failure cases to short-term success, while they remain highly accurate (84.41%) in identifying long-term success. In contrast, Qwen-3-30B and other language models primarily struggle to distinguish between short- and long-term success.

This pattern is consistent across experts, who substantially outperform language models in identifying long-term success (71.97% vs. 49.88%), though they are more susceptible to other types of errors. Overall, performance variance is considerably higher among language models (0.457–0.580) than among human experts (0.573–0.642).

Taken together, these results indicate that human experts remain more reliable and consistent, despite the near-expert performance achieved by the best-performing language model. Nevertheless, the complementary nature of errors made by humans and models suggests that current language models may already provide value as assistive tools rather than replacements for educational experts.

4 Conclusions

We introduce Edustories, a new collection of 1,492 authentic, teacher-written case studies from elementary and high-school classrooms. The dataset was proofread and anonymized through a two-stage process that included human review, making it, to our best knowledge, the first freely available educational dataset documenting classroom practices at this level of detail.

Edustories enables a wide range of research directions in education, linguistics, and AI, including sentiment analyses, identification of predictive features of successful interventions, and evaluation of AI assistants in collective educational practice. In this work, we focus on one such direction: assessing the ability of current LLMs in providing feedback on teachers' proposed interventions.

Our results show that while the best-performing models can reach the lower bound of human expert performance, language models remain less consistent and exhibit different failure modes than humans. These findings highlight the importance of developing specialized educational benchmarks before using AI systems in practice, but also underscore the existing potential of AI as assistants for educational practice.

Limitations

We wish to acknowledge several limitations of our work, possibly inviting future work towards addressing those. First, we note that the scope of our evaluations in outcome prediction does not fully reflect the complexity of the real-world deployment of assistive AI into the educational practice. While we focus on outcome prediction for simplicity, the feedback potentially accepted directly by the education practitioners would need to contain a free-form justification. The efficiency of such feedback is difficult to evaluate, but, we believe, presents a very interesting direction for future work.

Second, we acknowledge the limitation of Edustories dataset in the quality of English translation of free-form texts, constrained by the precision of the commercial translation services. While we pick the translation service for free-form texts systematically based on human assessment from among Google Translate, DeepL and Open-source Czech-English translation models, we note that errors may occus and estimate that 5–10% of stories may contain a type or "translationese" text. However, as we do not dispose of sufficient power of human proofreaders to manually verify all the English versions of our stories, we will appreciate the effort of the community in helping us to identify cases where the quality of free-form features of Edustories could be improved and possibly helping us to address them.

Finally, we acknowledge the limitation of the scope our evaluated models, that do not include the models over 100-billion parameters due to our computational restraints. Note that we intentionally exclude the proprietary API AI services from evaluations as these may not present an apples-to-apples comparison to raw language models evaluated locally. Furthermore, we believe that controlled, locally-deployable models correspond better with the practical application of feedback delivery systems as these systems will operate directly with sensitive data from classrooms that may not be safely shared with third parties.

References

Suyoung Bae, Gunhee Cho, Yun-Gyung Cheong, and Boyang Li. 2025. CharMoral: A character morality dataset for morally dynamic character analysis in long-form narratives. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8809–8818, Abu Dhabi, UAE. Association for Computational Linguistics.

Asta Batūraitė-Bunka, Liucija Masiūnienė, and Rita Žukauskienė. 2024. Unveiling the effects of non-violent communication training on youth empathy. *Social Welfare: Interdisciplinary Approach*, 14(2):xx–xx.

Catherine P. Bradshaw, Tracy E. Waasdorp, and Philip J. Leaf. 2012. Effects of school-wide positive behavioral interventions and supports on child behavior problems. *Pediatrics*, 130(5):e1136–e1145.

Ting Cai, Yu Xiong, Chengyang He, Chao Wu, and Linqin Cai. 2025. Classroom teacher behavior analysis: The tbu dataset and performance evaluation. *Computer Vision and Image Understanding*, 257:104376.

Christina Cipriano, Michael J. Strambler, Lauren H. Naples, Cheyeon Ha, Megan Kirk, Miranda Wood, Kaveri Sehgal, Almut K. Zieher, Abigail Eveleigh, Michael McCarthy, Melissa Funaro, Annett Ponnock, Jason C. Chow, and Joseph Durlak. 2023. The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based sel interventions. *Child Development*, 94(S1):1–24.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *ArXiv*, abs/2012.15738.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Trevor Fronius, Sean Darling-Hammond, Hannah Persson, Sarah Guckenburg, Nancy Hurley, and Anthony Petrosino. 2019. Restorative justice in u.s. schools: An updated research review. Technical report, WestEd Justice & Prevention Research Center, San Francisco, CA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and Amy Yang et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Florian Hofmann, Tina-Myrica Daunicht, Lea Plößl, and Michaela Gläser-Zikuda. 2025. Promoting reflection skills of pre-service teachers—the power of ai-generated feedback. *Education Sciences*, 15(10).

Chenglei Huang, Jia Zhu, Yilong Ji, Weijie Shi, Min Yang, Hanghui Guo, Jianxia Ling, Pasquale De Meo, Zilong Li, and Zhangze Chen. 2025. A multi-modal dataset for teacher behavior analysis in offline classrooms. *Scientific Data*, 12.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sean P. Kearney. 2019. The challenges of beginning teacher induction: a collective case study. *Teaching Education*, 32:142 – 158.

Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, and Christophe Gravier. 2025. Histoires morales: A french dataset for assessing moral alignment. In *North American Chapter of the Association for Computational Linguistics*.

Lihua Lin, Haodong Yang, Qingchuan Xu, Yanan Xue, and Dan Li. 2024. Research on student classroom behavior detection based on the real-time detection transformer algorithm. *Applied Sciences*, 14(14).

Tomás Matos, Walter Santos, Eftim Zdravevski, Paulo Jorge Coelho, Ivan Miguel Pires, and Filipe Madeira. 2025. A systematic review of artificial intelligence applications in education: Emerging trends and challenges. *Decision Analytics Journal*, 15:100571.

Jan Nehyba and Michal Štefánik. 2022. Applications of deep language models for reflective writings. *Education and Information Technologies*.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named entities in czech: Annotating data and developing NE tagger. In *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, number XVII in Lecture Notes in Computer Science, pages 188–195, Berlin / Heidelberg. Springer.

Michal Štefánik and Jan Nehyba. 2021. Czech and english reflective dataset (CEReD).

Auli Toom and Jukka Husu. 2021. *Classroom Interaction Challenges as Triggers for Improving Early Career Teachers' Pedagogical Understanding and Competencies Through Mentoring Dialogues*, pages 221–241. Springer International Publishing, Cham.

Iraklis Varlamis. 2025. Messy data in education: Enhancing data science literacy through real-world datasets in a master's program. *Education Sciences*, 15(4).

Walton Family Foundation. 2025. The AI dividend: New survey shows AI is helping teachers reclaim valuable time. `https://www.waltonfamilyfoundation.org/`. Article: "The AI Dividend." Accessed: 2026-01-05.

Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024a. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252:124167.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chengming Zhang. 2024. *Fostering Pre-Service Teacher Reflection through AI- Based Feedback: From Understanding AI Acceptance to Developing Effective AI-Driven Feedback*. Ph.D. thesis, Der Philosophischen Fakultät und Fachbereich Theologie der Friedrich-Alexander-Universität Erlangen-Nürnberg.

Chengming Zhang, Florian Hofmann, Lea Plößl, and Michaela Gläser-Zikuda. 2024. Classification of reflective writing: A comparative analysis with shallow machine learning and pre-trained language models. *Educ. Inf. Technol.*, 29(16):21593–21619.

A Experimental details

A.1 Confusion matrices

Tables 2 and 3 shows classification confusion matrix for the best-performing human annotator and best-performing language model, respectively. Tables 4 and 5 shows classification confusion matrix aggregated over all human experts, and evaluated language models, respectively.

	Predicted outcome		
True outcome	**fail**	**success-short**	**success-long**
fail	0.00	91.67	8.33
success-short	7.69	27.88	64.42
success-long	2.69	12.90	84.41

Table 2: **Best expert:** Row-normalized confusion matrix (%) for three-way outcome classification for the best-performing human expert.

	Predicted outcome		
True outcome	**fail**	**success-short**	**success-long**
fail	41.67	8.33	50.00
success-short	12.62	41.75	45.63
success-long	7.57	24.32	68.11

Table 3: **Best language model:** Row-normalized confusion matrix (%) for three-way outcome classification for the best-performing **language model** (Qwen3-30B).

	Predicted outcome		
True outcome	**fail**	**success-short**	**success-long**
fail	31.67	55.00	13.33
success-short	17.49	28.95	53.56
success-long	5.87	22.16	71.97

Table 4: **All experts:** Row-normalized confusion matrix (%) averaged across all human experts for three-way outcome classification.

	Predicted outcome		
True outcome	**fail**	**success-short**	**success-long**
fail	45.83	45.83	8.33
success-short	14.06	54.51	31.44
success-long	8.29	41.83	49.88

Table 5: **All language models:** Row-normalized confusion matrix (%) averaged across all language models for three-way outcome classification.

A.2 Evaluation Prompt

Table 6 displays the exact form of the best-performing prompt we applied in our evaluations. The selection process is detailed in Section 3.1.

Prompt

You will be given a short case study written by a teacher.
It describes a situation and the intervention chosen.

Your task is to classify the expected overall outcome of
the solution into exactly one category:
'Longterm success', 'Partial success', 'failure' or 'I don't know'

Important rules:
- Choose exactly one category.
- Reply with the category name ONLY, copied exactly.
- No punctuation, explanation, or extra text.

Case study:

"""{Description} {Anamnesis} {Intervention}"""

Response

One of {'Longterm success' / 'Partial success'/ 'failure' / 'I don't know'}

Table 6: Final prompt we use in our evaluations. Chosen as the best-performing variant among 5 candidates.

A.2.1 Evaluated models

In our experiments, we specifically evaluate the following models with HuggingFace identifiers: Qwen3-30B and Qwen3-4B introduced by Yang et al. (2025) corresponding to *Qwen3-30B-A3B-Instruct-2507* and *Qwen3-4B-Instruct-2507*, Llama 3.2-70B and Llama 3.3-3B introduced by (Grattafiori et al., 2024) corresponding to *meta-Llama-3.3-70B-Instruct* and *Llama-3.2-3B-Instruct*, Mistral-7B introduced by Jiang et al. (2023) corresponding to *Mistral-7B-Instruct-v0.3* and Olmo-3 introduced by Olmo et al. (2025) corresponding to *Olmo-3-7B-Instruct*.

B Use of AI Assistants

We acknowledge the use of AI assistants during the creation of this paper, namely in plots refinements, tables formatting, grammar control and text polishing. AI assistants were not employed in the creation of ideas, results and connections described in this paper.