Prompt-Driven Synthetic Data Generation for MITI Behavior Code Classification: A Diversity-Aware Few-Shot Approach

Gabriel Santiago Robles-Robles[1,*], Mario Alejandro Castro-Lerma[1], Jehu Jonathan Ramirez-Ramirez[1] and Jesus Antonio Flores-Briones[1]

[1]*Universidad de Sonora, Departamento de Matemáticas, Blvd. Luis Encinas y Rosales, Col. Centro, Hermosillo, Sonora 83000, Mexico*

Abstract

We present a synthetic data generation pipeline for training Motivational Interviewing (MI) behavior code classifiers, developed for the MIRROR@IberLEF 2026 shared task. The challenge requires participants to produce three balanced datasets of clinical conversation turns, each covering a pair of Motivational Interviewing Treatment Integrity (MITI 4.2.1) behavior codes: Simple vs. Complex Reflection (SR/CR), Open vs. Closed Question (OQ/CQ), and Giving Information vs. Persuasion (GI/PE). Our approach combines a fine-grained stratified diversity matrix with structured few-shot prompting over `Gemma 4 26B-A4B` loaded in 4-bit quantization, followed by a two-stage curation pipeline consisting of semantic deduplication via multilingual sentence embeddings and a format-based hard filter for question-type turns. The system ranked 1st overall among the participating teams (AVG Macro F1: 0.7366). Our highest individual score was on the question (OQ/CQ) pair (0.8559), where the field performed uniformly well; our margin over the second-ranked team came from a consistent advantage across all three pairs rather than a single dominant task. These results suggest that combining diversity stratification, code-contrastive prompting, and a single-question format constraint is an effective approach to generating MITI training data, while the finer-grained reflection (SR/CR) distinction remains more challenging.

Keywords

Motivational Interviewing, Synthetic Data Generation, MITI, Large Language Models, Few-Shot Prompting, Clinical NLP

1. Introduction

Motivational Interviewing (MI) is an evidence-based clinical communication style designed to strengthen a person's motivation and commitment to change [1]. Its effectiveness across substance use and a range of other health-related behaviors has driven interest in automated tools for assessing clinician adherence to MI principles [2]. The Motivational Interviewing Treatment Integrity (MITI) rubric [3] provides a standardized framework for coding clinician behavior, categorizing utterances into types such as reflections, questions, and information-giving behaviors, each with fine-grained distinctions that require both clinical expertise and contextual understanding to annotate reliably.

Developing automated MITI classifiers is constrained by data scarcity: expert-annotated therapy transcripts are scarce, and producing them requires costly annotation by coders with specialized training, on top of the difficulty of sharing real clinical recordings [4, 5]. Synthetic data generation using Large Language Models (LLMs) has emerged as a promising approach to address this bottleneck, as modern instruction-tuned models can produce linguistically fluent and contextually appropriate text across diverse scenarios [6, 7].

However, in our preliminary experiments naïve LLM generation exhibited recurring failure modes: repetitive phrasing, distribution collapse onto dominant demographic profiles, and blurring of subtle

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

gsrrobles@gmail.com (G. S. Robles-Robles); mar.castro.lerma@gmail.com (M. A. Castro-Lerma); jonathanjehu@gmail.com (J. J. Ramirez-Ramirez); jesusantoniofloresbriones@gmail.com (J. A. Flores-Briones)

0009-0008-2773-6567 (G. S. Robles-Robles); 0009-0007-1026-4498 (M. A. Castro-Lerma); 0009-0005-3890-7497 (J. J. Ramirez-Ramirez); 0009-0004-3919-5412 (J. A. Flores-Briones)

code boundaries. The SR/CR distinction (paraphrasing without inference vs. reflecting with added meaning) and the OQ/CQ distinction (open vs. closed questions) are susceptible to boundary confusion in generated data [8].

This paper describes the system developed by team UNISON-GOATS for the MIRROR@IberLEF 2026 Challenge [9], part of the IberLEF 2026 evaluation campaign [10], which placed **1st overall** (AVG F1: 0.7366). Our key contributions are:

1. A **stratified diversity matrix** covering six clinical and demographic axes, designed to reduce the risk of distribution collapse onto narrow topic or demographic niches.
2. **Code-contrastive prompting** that provides both the target code definition and its paired contrast code, forcing the model to generate unambiguously boundary-clear examples.
3. A **two-stage curation pipeline** combining semantic deduplication with a format-based hard filter specific to question-type turns.

2. Related Work

Researchers have pursued automated recognition of MI-adherent clinician behavior as a tool for training feedback and treatment quality monitoring [2, 11]. Early work relied on manually engineered features and statistical models; subsequent approaches shifted to machine learning and neural networks, with systems modeling MI behavior codes directly from session transcripts [12, 13].

The scarcity of annotated MI transcripts has motivated several data augmentation strategies. Token-level edit operations [14] and back-translation [15] increase lexical diversity while preserving label integrity. More recently, LLM-based generation has been explored as a means of producing task-specific training data from scratch, with prompt engineering playing a central role in controlling output quality and adherence to target label definitions [16, 17, 18].

Few-shot prompting improves performance on classification tasks; in-context example selection and quality have a substantial effect on output [6, 19, 20]. Contrastive demonstrations (positive examples of a target class alongside counter-examples) sharpen a model's discrimination between adjacent categories [21].

Instruction-tuned models from the Gemma family [22] have demonstrated strong multilingual performance while remaining computationally accessible through 4-bit quantization [23]. Mixture-of-Experts variants further lower inference cost by activating only a fraction of their parameters per token [24], which makes them suitable for generation tasks requiring long prompts with multiple in-context examples. Sentence-embedding semantic deduplication reduces redundancy in large text corpora without discarding topically distinct samples [25].

3. Methods

The MIRROR task requires generating three balanced CSV datasets, one per MITI behavior code pair, each containing between 100 and 200 labeled conversation turns. Each row represents a single conversation turn with three fields: `client_utterance`, `clinician_utterance`, and `behavior_code_1`. The shared-task platform (https://mirror-iberlef.vercel.app/) concatenates the client and clinician utterances and uses them to fine-tune a frozen DistilBERT model via LoRA on the submitted data; the resulting model is then evaluated against a hidden gold standard of real clinical interactions using Macro F1 score. Figure 1 gives an overview of the full system, from the seed data and diversity matrix through code-contrastive generation and two-stage curation to the platform's evaluation.

3.1. Seed Data and Task Structure

Three seed datasets were provided by the organizers, each containing 200 rows (100 per class):

- `INFO_AUG.csv`: codes GI and PE.

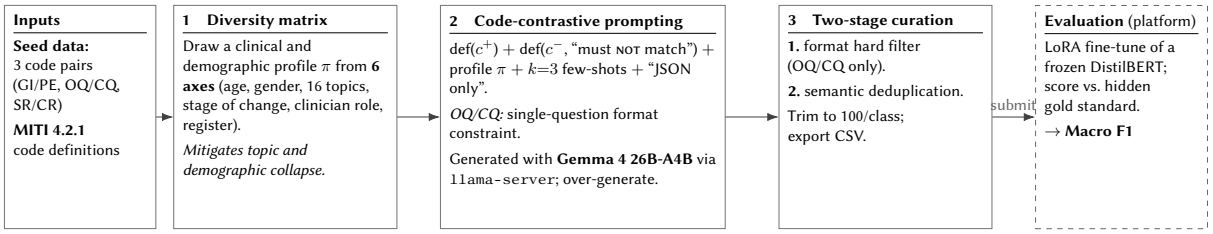

Figure 1: The three components of our method. From the seed data and the MITI 4.2.1 code definitions, (1) a **stratified diversity matrix** samples a clinical and demographic profile π per example to counter distribution collapse; (2) **code-contrastive prompting** pairs the target-code definition c^+ with its contrast c^- ("must not match"), adds the profile, $k{=}3$ few-shot examples, and (for the OQ/CQ pair) a single-question format constraint, generating each turn with `Gemma 4 26B-A4B` through `llama-server`; and (3) a **two-stage curation** discards malformed question turns and removes semantic near-duplicates. The curated datasets are submitted to the shared-task platform, which fine-tunes a frozen DistilBERT via LoRA and scores it against a hidden gold standard (external).

- `QUEST_AUG.csv`: codes OQ and CQ.
- `REFLEX_AUG.csv`: codes SR and CR.

Each dataset contains `client_utterance`, `clinician_utterance`, and `behavior_code_1`. Client utterances repeat across rows (the same patient turn is paired with two distinct clinician responses, one per code). The seed data covers a narrow range of topics dominated by substance use and clinical supervision, making topical expansion a key challenge for the generation pipeline.

3.2. MITI Code Definitions

Each generation prompt is grounded in precise MITI 4.2.1 operational definitions:

- **SR** (*Simple Reflection*): the clinician repeats or paraphrases the patient's words without adding inference, emotion, or emphasis.
- **CR** (*Complex Reflection*): the clinician adds meaning, infers emotion, identifies ambivalence, or reframes using metaphor, going beyond the literal content of the patient's statement.
- **OQ** (*Open Question*): a question that cannot be answered with yes/no or a single datum; invites elaboration. Constrained to begin with ¿ and contain exactly one question.
- **CQ** (*Closed Question*): a question answerable with yes/no, a specific option, or a concrete datum. Same format constraints as OQ.
- **GI** (*Giving Information*): the clinician shares facts, explains a mechanism, or describes a resource neutrally, without attempting to change the patient's behavior.
- **PE** (*Persuasion*): the clinician attempts to change the patient's opinion or behavior through argument, warning, or unsolicited advice, not fully respecting patient autonomy.

3.3. Stratified Diversity Matrix

To prevent the distribution collapse observed with a preliminary generator (`gpt2-small-spanish`), which collapsed onto tobacco and marijuana topics and produced repetitive sentence starters, we define a six-axis diversity matrix $\mathcal{D}$, shown in Table 1. For each sample, a demographic and clinical profile π_i is drawn uniformly at random from $\mathcal{D}$ and embedded in the generation prompt as contextual inspiration, without being quoted literally in the generated text.

The register axis was fixed to *neutral Spanish* in the final system. Preliminary experiments that included regional variants (Mexican colloquial, Rioplatense, Peninsular formal, Andean colloquial) produced stylistic drift inconsistent with the format and register of the seed data, reducing alignment between generated samples and the classifier's training distribution.

Table 1
Stratified diversity matrix $\mathcal{D}$ used for profile sampling.

Axis	Values
Age	18–25, 26–40, 41–60, 60+
Gender	woman, man, non-binary
Clinical topic	16 topics (tobacco, alcohol, marijuana, cocaine, anxiety, depression, chronic pain, insomnia, eating disorder, sedentary lifestyle/obesity, type 2 diabetes, hypertension, antiretroviral adherence, pathological gambling, domestic violence (victim), unresolved grief)
Stage of change	precontemplation, contemplation, preparation, action, maintenance, relapse
Clinician role	family physician, clinical psychologist, primary care nurse, social worker, psychiatrist, nutritionist
Language register	neutral Spanish

3.4. Generation Model

We use `Gemma 4 26B-A4B-it` [22], an instruction-tuned, decoder-only Mixture-of-Experts (MoE) transformer with approximately 25.2 billion total parameters, of which only about 3.8 billion are active per token, and with multilingual support that includes Spanish. The model is run from the Unsloth 4-bit quantized GGUF checkpoint `gemma-4-26B-A4B-it-UD-Q4_K_XL`[1] served locally through `llama.cpp` (`llama-server`); the generation pipeline issues requests to this local server and collects the returned completions. The 17 GB quantized checkpoint is loaded entirely into the 32 GB of VRAM of an NVIDIA RTX 5090, leaving ample margin for the embedding model used during curation. No fine-tuning is performed; all task conditioning is achieved through prompt engineering.

We use the sampling configuration recommended for Gemma: temperature $= 1.0$, top-$p = 0.95$, and top-$k = 64$, with a maximum of 4096 generated tokens per request. No random seed was fixed, so the generation pipeline is not deterministic across runs.

3.5. Code-Contrastive Few-Shot Prompting

For each sample targeting code c^+ with contrast code c^-, the generation procedure is as follows:

1. Draw a diversity profile π_i from $\mathcal{D}$.
2. Sample $k = 3$ in-context examples from the seed dataset for class c^+.
3. Build the prompt with: (a) a system persona as a MITI expert; (b) the operational definition of c^+ (Section 3.2); (c) the operational definition of c^- with an explicit instruction that the output must *not* match it; (d) the profile π_i as contextual inspiration; (e) the k few-shot examples; and (f) a strict instruction to return a valid JSON object with keys `client_utterance` and `clinician_utterance` and no additional text.

The contrastive component is the structural core of our prompting strategy. By explicitly defining the boundary between c^+ and c^- within the same prompt, the model is required to generate samples that are unambiguously on the correct side of the decision boundary. This matters most for the SR/CR pair, where the distinction between a paraphrase and an interpretive reflection blurs without an explicit negative constraint.

For the OQ/CQ pair, an additional constraint is appended to the prompt: the generated `clinician_utterance` *must begin with ¿ and contain exactly one question.* This constraint was

[1] https://huggingface.co/unsloth/gemma-4-26B-A4B-it-GGUF/blob/main/gemma-4-26B-A4B-it-UD-Q4_K_XL.gguf

System

Eres un experto clínico en Entrevista Motivacional (MI) y en la rúbrica MITI 4.2.1. Tu trabajo es generar turnos de conversación sintéticos en español, realistas y clínicamente plausibles, para entrenar clasificadores de códigos de comportamiento. Respetas siempre la definición exacta del código objetivo. Devuelves SOLO un JSON válido, sin texto adicional, sin markdown, sin comentarios.

User (instantiated for the OQ/CQ pair; target $c^+ =$ OQ, contrast $c^- =$ CQ)

Genera UN turno de conversación cliente-clínico que ejemplifique de forma INEQUÍVOCA el código **OQ**.

Definición del código objetivo (OQ): pregunta abierta que no se puede responder con sí/no ni con un dato corto; invita a elaborar, explorar, narrar; suele empezar con "qué", "cómo", "de qué manera".

Para evitar confusión, NO debe encajar en este otro código del par (CQ): pregunta cerrada que se responde con sí/no, una opción de un conjunto pequeño, o un dato concreto; limita la respuesta.

Perfil del paciente y contexto clínico (úsalo como inspiración, no lo cites literalmente):
edad: 26–40; género: mujer; tema: consumo de alcohol; etapa de cambio: contemplación; profesional: psicólogo clínico; registro: español neutro.

Ejemplos de referencia (mismo código objetivo, $k = 3$ del seed):
`Ejemplo 1: {"client_utterance": ". . . ", "clinician_utterance": ". . . "}` [+2 more]

Restricciones: `client_utterance` de 1–3 oraciones, voz del paciente; `clinician_utterance` de 1–2 oraciones, debe ejemplificar OQ estrictamente; **debe empezar con ¿ y contener exactamente una pregunta**; evita copiar frases de los ejemplos; no incluyas el nombre del código en el texto.

Devuelve SOLO el JSON con las claves `client_utterance` y `clinician_utterance`.

Figure 2: Code-contrastive prompt instantiated for the OQ/CQ pair (target code $c^+ =$ OQ). The prompt is issued in Spanish, the language of the generated data. It shows the system persona, the paired c^+/c^- definitions, the sampled diversity profile, the $k = 3$ seed few-shots, and the JSON-only instruction; the single-question format constraint specific to the OQ/CQ pair is highlighted in bold. Definitions, constraints and few-shot bodies are abbreviated for space.

motivated by systematic inspection of generation failures in a preliminary pipeline, where the model produced mixed-mode turns such as:

"Entiendo que estás pasando por un mal momento, ¿Has intentado ir con un especialista?"

These hybrid empathy-plus-question structures are clinically natural but diverge from the single-question format of the seed data, causing misclassification by the downstream frozen classifier. Figure 2 shows the complete prompt instantiated for the OQ/CQ pair, including this format constraint.

3.6. Post-Generation Curation

Generation proceeds by oversampling: we target 130 samples per class (30% above the final target of 100) to provide curation headroom. Curation proceeds in two stages.

Stage 1: Format hard filter (OQ/CQ only). Any generated `clinician_utterance` that does not begin with '¿' is discarded immediately, before computing any embeddings. This is an $O(1)$ string check that removes the most common generation failure mode for question turns at negligible computational cost.

Stage 2: Semantic deduplication. For each class independently, all retained turns (client + clinician utterance concatenated) are encoded using `paraphrase-multilingual-MiniLM-L12-v2` [26]. Samples are processed sequentially; a new sample is discarded if its cosine similarity to any already-kept sample exceeds the threshold $\tau = 0.92$. This prevents lexical near-duplicates from inflating the dataset size without contributing distributional diversity. After deduplication, each class is trimmed to 100 samples and the final dataset is shuffled before export.

Table 2
Macro F1 per task for all participating teams in the MIRROR challenge, ordered by overall average (UNISON-GOATS is our system). Results are averages over 5 runs; standard deviations in parentheses. Best score per column in bold.

Team	Rank	Information	Questions	Reflections	AVG
UNISON-GOATS	1	0.7728 (0.0066)	0.8559 (0.0042)	0.5811 (0.0175)	**0.7366 (0.0064)**
OpenCodice	2	0.7466 (0.0183)	**0.8663 (0.0045)**	0.5942 (0.0140)	0.7357 (0.0061)
umuteam	3	0.7593 (0.0039)	0.8464 (0.0027)	0.5875 (0.0159)	0.7311 (0.0054)
ucdicsnlpteam	4	0.7603 (0.0112)	0.8438 (0.0047)	0.5869 (0.0153)	0.7303 (0.0062)
steam_uaz	5	**0.7749 (0.0055)**	0.8382 (0.0052)	0.5667 (0.0131)	0.7266 (0.0056)
roman_lab	7	0.7369 (0.0135)	0.8311 (0.0066)	**0.5948 (0.0168)**	0.7209 (0.0068)

4. Experiments

All generation and curation was performed on a workstation equipped with an NVIDIA RTX 5090 GPU (32 GB VRAM). Text generation was served by a local `llama.cpp` instance (`llama-server`) hosting the quantized model, while semantic deduplication used the `sentence-transformers` library for embedding computation. Generating the oversampled candidates for each dataset pair took approximately 18 seconds.

The final submitted datasets contain 200 rows each (100 per class). CSV files were exported with UTF-8-BOM encoding to ensure correct rendering of Spanish diacritics. Submitted datasets were validated automatically by the platform for row-count constraints before triggering the LoRA fine-tuning and evaluation pipeline.

5. Results and Discussion

Table 2 reports corrected results averaged over 5 runs. Our system ranked **1st overall** with an average Macro F1 of 0.7366 ($\pm$0.0064), ahead of OpenCodice (0.7357) by a narrow margin of 0.0009.

The corrected evaluation reveals a more uniform picture across the field than preliminary scores suggested. On the Questions (OQ/CQ) pair, performance was uniformly high (range 0.83–0.87): our score of 0.8559 was competitive but slightly below the best (OpenCodice, 0.8663), suggesting that the single-question format constraint contributed to strong question generation without being uniquely superior. On Information (GI/PE), scores clustered between 0.74 and 0.77, with steam_uaz obtaining the highest score (0.7749) and our system second (0.7728). The Reflections (SR/CR) pair showed the most variance across teams (0.57–0.59): roman_lab obtained the best individual score (0.5948), while our system scored 0.5811, consistent with SR/CR remaining the hardest pair across all systems.

Our first-place ranking therefore reflects consistent performance across all three pairs rather than a dominant score on any single task. The narrow gap to OpenCodice suggests that the field converged on broadly similar quality levels, and that our combination of diversity stratification, code-contrastive prompting, and two-stage curation produced well-rounded synthetic datasets without a critical weakness on any pair. The SR/CR distinction depends on whether the clinician adds interpretive depth, a boundary operationalized in the MITI coding manual [8] but recognized as difficult to code reliably, which makes it difficult to generate and evaluate consistently across all systems.

6. Conclusions

We presented a prompt-driven synthetic data generation system for MITI behavior code classification that ranked 1st overall in the MIRROR@IberLEF 2026 challenge (AVG Macro F1: 0.7366). Our approach combines a stratified diversity matrix intended to reduce topic and demographic collapse, code-contrastive few-shot prompting aimed at sharpening code boundaries in generated samples, and a two-stage curation pipeline of format-based hard filtering and semantic deduplication. Across the

shared task, our first-place ranking reflects consistent performance across all three pairs rather than a dominant score on any single task.

On Information/GI/PE (0.7728) scores were tightly clustered across teams, consistent with these two codes being the most semantically separable. On Questions/OQ/CQ, the field performed uniformly well (0.83–0.87 range), with our score of 0.8559 competitive but not uniquely superior, suggesting the format constraint helped but was not the sole differentiator. SR/CR remained the hardest pair (0.5811): generating data that reliably teaches a classifier this subtle linguistic boundary requires either more targeted prompt engineering or post-hoc verification of generated samples against the MITI rubric.

Acknowledgments

The authors acknowledge the support of the High-Performance Computing Area of the Universidad de Sonora (ACARUS) for the use of its supercomputing infrastructure, which was fundamental to the development of this work (https://acarus.unison.mx), as well as the financial support provided by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through a graduate scholarship.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: Grammar and spelling check, Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. R. Miller, S. Rollnick, Motivational Interviewing: Helping People Change, 3rd ed., Guilford Press, New York, 2013.

[2] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, V. Srikumar, A comparison of natural language processing methods for automated coding of motivational interviewing, Journal of Substance Abuse Treatment 65 (2016) 43–50.

[3] T. B. Moyers, T. Martin, J. K. Manuel, S. M. Hendrickson, W. R. Miller, Assessing competence in the use of motivational interviewing, Journal of Substance Abuse Treatment 28 (2005) 19–26.

[4] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. Reforgiato Recupero, D. Riboni, Anno-MI: A dataset of expert-annotated counselling dialogues, in: ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2022, pp. 6177–6181.

[5] A. Welivita, P. Pu, Curating a large-scale motivational interviewing dataset using peer support forums, in: Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022), 2022, pp. 3315–3330.

[6] T. Brown, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33, 2020, pp. 1877–1901.

[7] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, H. Wang, On LLMs-driven synthetic data generation, curation, and evaluation: A survey, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024. ArXiv:2406.15126.

[8] T. B. Moyers, T. Martin, J. K. Manuel, W. R. Miller, D. Ernst, Revised global scales: Motivational interviewing treatment integrity 3.1.1, 2010. Unpublished manuscript, University of New Mexico.

[9] L. J. Arellano, C. Olachea, J. Piette, H. J. Escalante, L. Villaseñor-Pineda, M. M. y Gómez, D. I. Hernández-Farías, Overview of MIRROR at IberLEF 2026: Motivational interviewing response & rating via synthetic conversational turns, Procesamiento del Lenguaje Natural 77 (2026).

[10] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian

Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[11] N. Flemotomos, V. R. Martinez, Z. Chen, K. Singla, V. Ardulov, R. Peri, D. D. Caperton, J. Gibson, M. J. Tanana, P. Georgiou, J. Van Epps, S. P. Lord, T. Hirsch, Z. E. Imel, D. C. Atkins, S. Narayanan, Automated evaluation of psychotherapy skills using speech and language technologies, Behavior Research Methods 54 (2022) 690–711.

[12] D. Can, R. A. Marín, P. G. Georgiou, Z. E. Imel, D. C. Atkins, S. S. Narayanan, "it sounds like…": A natural language processing approach to detecting counselor reflections in motivational interviewing, Journal of Counseling Psychology 63 (2016) 343–350.

[13] J. Cao, M. Tanana, Z. E. Imel, E. Poitras, D. C. Atkins, V. Srikumar, Observing dialogue in therapy: Categorizing and forecasting behavioral codes, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 2019, pp. 5599–5611.

[14] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: Proceedings of EMNLP-IJCNLP 2019, 2019, pp. 6382–6388.

[15] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 2018, pp. 489–500.

[16] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), 2023, pp. 13484–13508.

[17] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating training data with language models: Towards zero-shot language understanding, in: Advances in Neural Information Processing Systems 35, 2022.

[18] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, L. Kong, ZeroGen: Efficient zero-shot learning via dataset generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), 2022, pp. 11653–11669.

[19] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for GPT-3?, in: Proceedings of Deep Learning Inside Out (DeeLIO 2022), 2022, pp. 100–114.

[20] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021), 2021, pp. 3816–3830.

[21] Y. Mo, J. Yang, J. Liu, S. Zhang, J. Wang, Z. Li, C-ICL: Contrastive in-context learning for information extraction, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024. ArXiv:2402.11254.

[22] Gemma Team, Google DeepMind, Gemma 4 model card, Google AI for Developers, 2026. URL: https://ai.google.dev/gemma/docs/core/model_card_4.

[23] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems 36, 2023.

[24] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, et al., Mixtral of experts, 2024. ArXiv:2401.04088.

[25] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, A. S. Morcos, SemDeDup: Data-efficient learning at web-scale through semantic deduplication, 2023. ArXiv:2303.09540.

[26] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of EMNLP 2020, 2020.