Divide, Cache, Conquer

Dichotomic Prompting for Efficient Multi-Label LLM-Based Classification

Mikołaj Langner, Jan Eliasz, Ewa Rudnicka, Jan Kocoń

Department of Artificial Intelligence, Wroclaw Tech, Poland

Why Affective Analysis is Difficult

Subjectivity

Emotions are context-dependent. Is "crying" sad or joyful? This requires subtle understanding that simple keywords miss.

Concept Drift

Language evolves rapidly. New labels (e.g., "Cringe", "Hype") emerge, making static taxonomies obsolete.

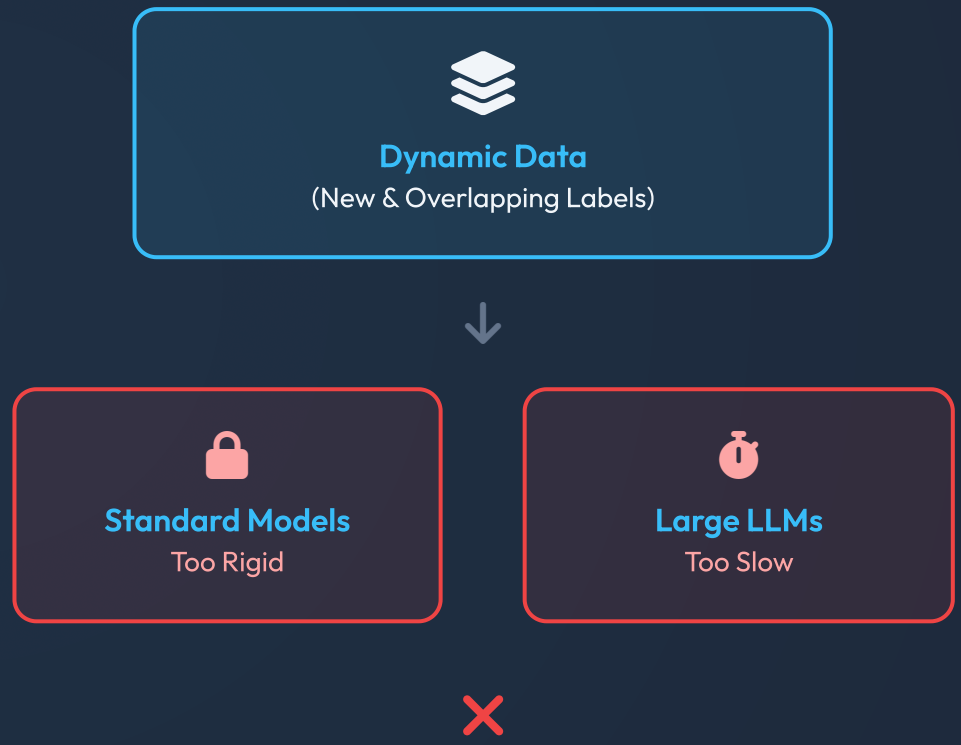
High Overlap

Affective states are not mutually exclusive. A text can be "Ironic", "Funny", and "Offensive" simultaneously.

The Architectural Dilemma

We are currently facing a trade-off between **capability** and **cost**.

- **Rigid Tools:** Traditional encoders (BERT) are fast but require expensive retraining for any new label.
- **Costly Flexibility:** Modern LLMs offer Zero-Shot adaptation but are too slow for real-time scale.



Standard Approach: Structured JSON

The conventional way to use LLMs for this task is to ask for a **single structured output**.

- **High Complexity:** The model must track all 24 labels simultaneously.
- **Fragile:** A single syntax error in the JSON invalidates the entire response.
- **Slow Generation:** Generating a long JSON string token-by-token increases latency.

Our Approach: Dichotomic Prompting

We propose decomposing the task into **K independent binary decisions**.

- **Simplicity:** The model only answers "Yes" or "No".
- **Robustness:** No complex syntax to break. Can handle any number of labels dynamically.
- **Parallelizable:** Each label is queried independently.

Powered by Prefix Caching

- **Redundancy:** The Instruction and Input Text are identical for all 24 queries.
- **Optimized "Prefill":** The model processes this shared prefix *once* and stores the attention states in memory (KV Cache).
- **Fast Decoding:** For each label, only the tiny, unique question suffix ("Is it Joy?") is computed.
- **Result:** Processing 24 labels costs marginally more than processing just one.

Dataset Composition

Usage Contexts & Sources

A corpus of **10,000 Polish texts** aggregated from six distinct sources to ensure diversity.

CONTEXT	SOURCE CORPORA	COUNT
News	Wikinews PL	4,633
# Social Media	Twitter (713) + CDT (1,628)	2,341
Reviews	Allegro Reviews	2,000
Academic	Open Coursebooks	1,026

5 Balanced Topics

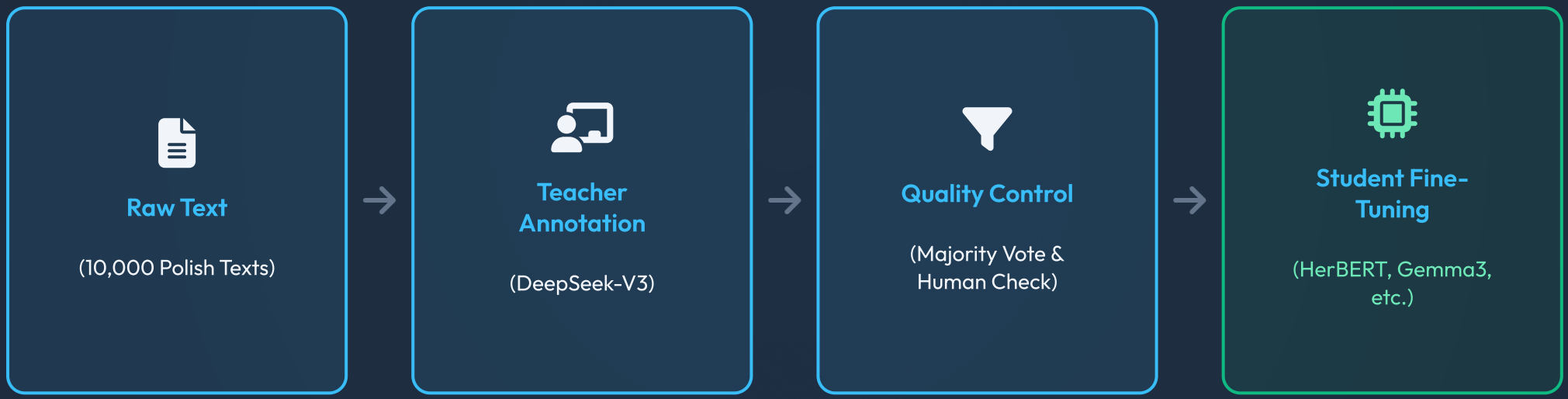
Politics, Sport, Science, Products, Culture

~2,000 texts per topic

24 Affective Labels

Comprehensive annotation schema covering emotions, sentiment, and specialized affective states.

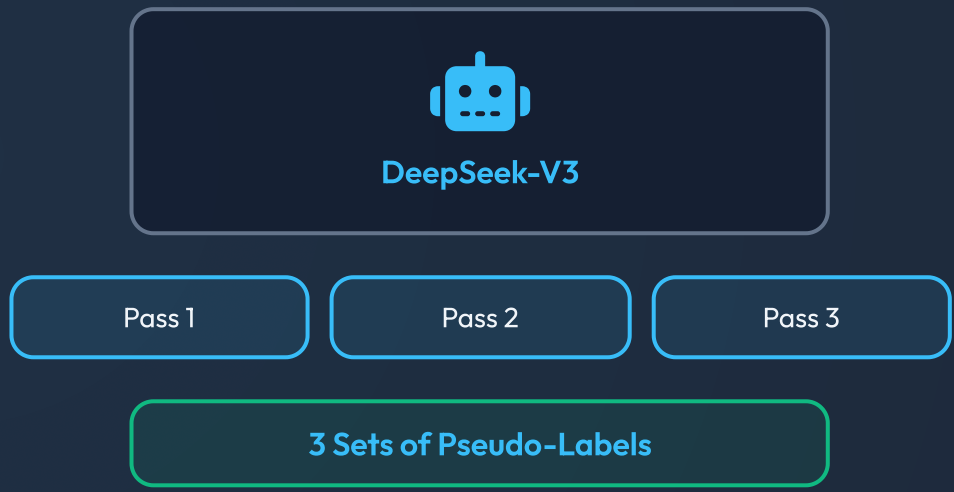
Methodology: The Distillation Pipeline



Step 1: Teacher Annotation

We leverage a massive, state-of-the-art LLM to generate our training data.

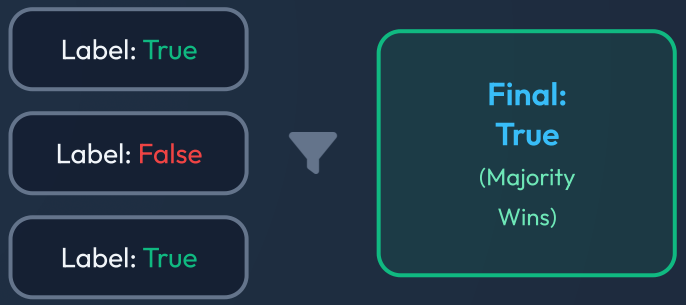
- **Teacher Model:** DeepSeek-V3 (Mixture-of-Experts).
- **Task:** Annotate each text for 24 affective dimensions.
- **Redundancy Strategy:** To ensure reliability and filter out hallucinations, the model generates **3 independent annotations** for every single text sample.



Step 2: Aggregation & Quality Control

Raw LLM outputs can be noisy. We rigorously filter them to create a "Silver Standard" dataset.

- **Majority Vote:** A label is assigned only if the Teacher model selected it in at least 2 out of 3 passes.
- **Human Verification:** A subset of data was manually verified by human experts to measure alignment.
- **Result:** High consistency (PSA > 0.8) between the aggregated LLM labels and human judgment.



Step 3: Student Fine-Tuning

The Students

We train much smaller, faster models to replicate the Teacher's performance. Models include **HerBERT** (Encoder) and **Gemma3-1B** (Decoder).

The Exam

Models are fine-tuned using **Binary Cross-Entropy Loss**. They learn to independently predict the probability of each of the 24 affective labels.

The Goal

To achieve **LLM-level accuracy** with **SLM-level speed** and deployment costs.

Evaluation Protocols

In-Distribution (ID)

The standard supervised learning setup.

- **Train:** All 24 Labels
- **Test:** All 24 Labels

Goal: Measure how well the model learns specific patterns seen during training.

Out-of-Distribution (OOD)

A challenging "Leave-One-Out" scenario.

- **Train:** 23 Labels (Label X Hidden)
- **Test:** Only Label X

Goal: Measure true *Zero-Shot Generalization*. Can the model understand a concept just from its name/prompt?

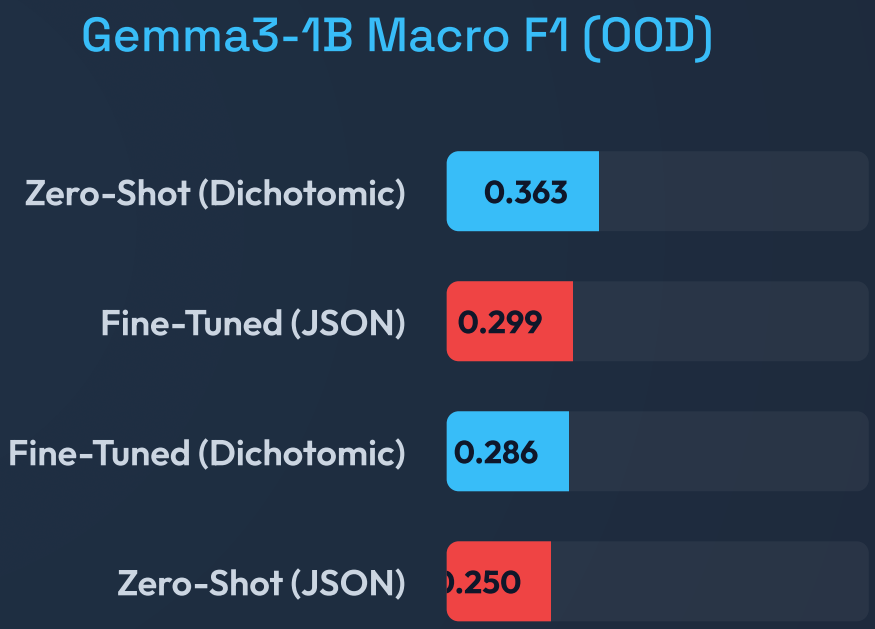
1. In-Distribution Results (Fine-Tuned)

Key Finding: *Dichotomic prompting (Blue) achieves comparable accuracy to complex JSON output (Red). Small Decoder models perform on par with the traditional Encoder Baseline (Green).*

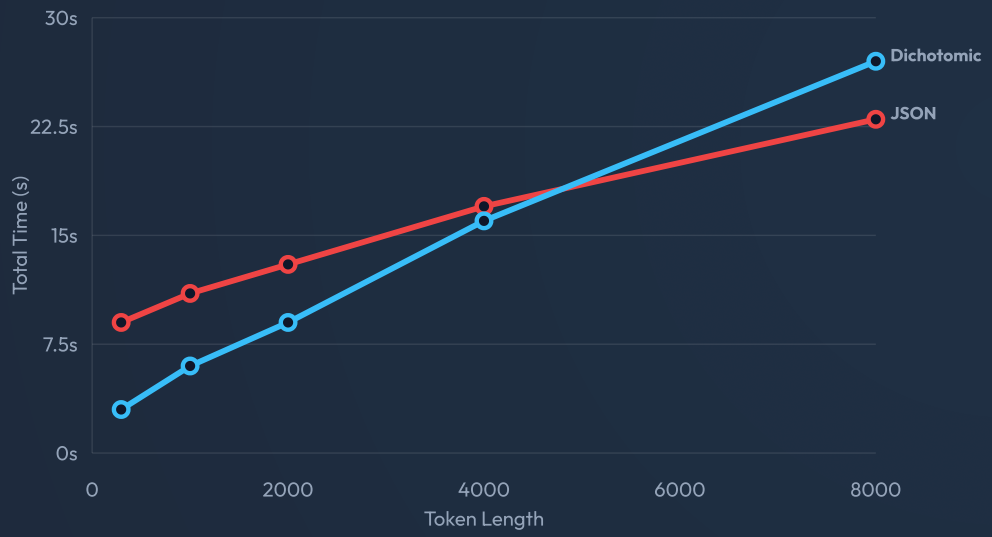
2. Out-of-Distribution Generalization

Scenario: The model encounters a label (e.g., "Ironic") it never saw during fine-tuning.

- **Winner:** Zero-Shot Dichotomic. The natural language question "Is it Ironic?" triggers the model's pre-trained knowledge.
- **Loser:** Zero-Shot JSON. The complex schema confuses the model without training.
- **The "Fine-Tuning Trap":** Fine-tuning improves seen labels but degrades performance on unseen ones (Overfitting).



Efficiency: Inference Time



Analysis (1k Texts)

Short Texts (< 4k tokens):

Dichotomic wins (3s vs 9s). Prefix caching optimizes the 24 small queries.

Crossover (≈ 4k tokens):

Performance equalizes (16s vs 17s) as repeated query overhead grows.

Long Texts (> 4k tokens):

JSON wins (23s vs 27s). Single-pass generation scales better here.

Verdict: Dichotomic dominates for typical inputs.

Conclusion: Divide, Cache, Conquer

By combining **Dichotomic Prompting** with **Prefix Caching**, we achieve the best of both worlds:

Flexible
Adapts to new labels (OOD).

Efficient
Fast inference on short texts.

Scalable
Enables distillation to SLMs.

Q&A

Thank you for your attention.