Divide, Cache, Conquer: Dichotomic Prompting for Efficient Multi-Label LLM-Based Classification

Mikołaj Langner, Jan Eliasz, Ewa Rudnicka, and Jan Kocoń
Department of Artificial Intelligence, Wroclaw Tech, Poland
{mikolaj.langner, jan.eliasz, ewa.rudnicka, jan.kocon}@pwr.edu.pl

***Abstract*—We introduce a method for efficient multi-label text classification with large language models (LLMs), built on reformulating classification tasks as sequences of dichotomic (yes/no) decisions. Instead of generating all labels in a single structured response, each target dimension is queried independently, which, combined with a prefix caching mechanism, yields substantial efficiency gains for short-text inference without loss of accuracy. To demonstrate the approach, we focus on affective text analysis, covering 24 dimensions including emotions and sentiment. Using LLM-to-SLM distillation, a powerful annotator model (DeepSeek-V3) provides multiple annotations per text, which are aggregated to fine-tune smaller models (HerBERT-Large, CLARIN-1B, PLLuM-8B, Gemma3-1B). The fine-tuned models show significant improvements over zero-shot baselines, particularly on the dimensions seen during training. Our findings suggest that decomposing multi-label classification into dichotomic queries, combined with distillation and cache-aware inference, offers a scalable and effective framework for LLM-based classification. While we validate the method on affective states, the approach is general and applicable across domains.**

***Index Terms*—Large Language Models, Text Classification, Multi-Label Classification, Dichotomic Prompting, Caching, Distillation, Affective Computing**

I. Introduction

Affective computing, first articulated by Picard [1], has grown into a vibrant research field with applications in sentiment analysis, social media monitoring, and human–computer interaction. In the textual domain, annotating data for affective dimensions has proven useful across a wide range of tasks, yet it remains challenging due to the overlapping, context-dependent, and evolving nature of affective states [2]–[20].

A significant advancement would be the development of a universal classifier capable of detecting an arbitrary set of affective classes without retraining. Traditional classifiers with a fixed K-class output are unsuitable for this purpose, as they offer neither flexibility in defining new classes nor customization of classification guidelines. Recent large language models (LLMs) have demonstrated strong zero-shot and few-shot capabilities [21], [22], making them well-suited for such dynamic settings. However, directly deploying powerful LLMs can be prohibitively expensive, motivating the use of smaller language models (SLMs) trained via distillation.

To address these challenges, we propose **dichotomic prompting**, which reformulates multi-label classification into independent yes/no queries and leverages prefix caching for efficiency. This strategy provides robustness in zero-shot settings while maintaining scalability.

In summary, this paper makes the following contributions:

- We introduce dichotomic prompting as an efficient mechanism for multi-label affective classification, yielding substantial inference gains for short texts when combined with prefix caching.
- We show that dichotomic prompting is particularly advantageous in zero-shot scenarios, mitigating errors from structured JSON generation and parsing.
- We empirically validate the approach on affective text analysis, demonstrating comparable or superior accuracy to structured prompting, with improved robustness and reduced computational cost.

II. Background and Motivation

A. Limitations of Traditional Classification Architectures

In affective computing, multi-label classification faces complexity due to overlapping, evolving, and context-sensitive labels. Traditional encoder-based models with fixed output heads (such as BERT [23]) are poorly suited for such settings. Extending their label space often demands retraining and architectural redesign, limiting adaptability and increasing computational burden.

This challenge has been long recognized in the broader multi-label learning literature. Early work formalized multi-label classification as a distinct problem setting [24], and surveys have emphasized the limitations of fixed-output models when label spaces evolve [25], [26]. More recently, research on dynamic label taxonomies and label evolution has highlighted the same tension in modern architectures, where incorporating new or subjective categories often requires retraining from scratch [27], [28]. Encoder–decoder models like T5 [29] introduce partial flexibility, but at a cost: their increased inference complexity presents practical limitations on deployment [30].

B. Decoder-Only LLMs for Prompt-Based Classification

Decoder-only language models offer a more adaptable alternative. Label prediction can be formulated as answering natural-language queries rather than requiring fixed-size outputs. This makes the approach inherently flexible: new labels can be introduced simply by crafting new prompts, without retraining the model.

This idea builds on the success of prompt-based and instruction-tuned LLMs, which demonstrated strong zero-shot

Acknowledgement

and few-shot capabilities [21], [22]. Systematic surveys of prompting methods [31] further underscore how query-based reformulations enable robust generalization across subjective or sparsely represented dimensions. Consequently, decoder-only models are particularly well-suited to affective computing, where categories are often subjective and prone to overlap. Indeed, prior comparative reviews suggest that such models generalize more effectively than conventional supervised classifiers in low-resource and domain-adaptive settings [32].

C. Prefix Caching for Efficient Prompt Execution

Decoder-only transformer architectures commonly implement *key–value (KV) caching* to optimize the computational efficiency of autoregressive generation. In this mechanism, the attention-related intermediate representations, specifically, the keys and values computed at each decoding step, are stored and reused in subsequent steps. This reuse obviates the need for recomputing these components during each forward pass, leading to substantial reductions in latency and compute overhead during inference [33].

Building upon this foundation, inference-optimized frameworks such as **vLLM** introduce the concept of *prefix caching*. This technique generalizes KV caching to enable reuse not only within a single decoding pass but across multiple prompts that share an identical input prefix [34]. Prefix caching is particularly advantageous in settings characterized by high prompt redundancy, such as multi-label classification with prompt-based formulations. In such cases, distinct classification queries often share a large portion of their input prompt, differing only in task-specific suffixes (e.g., candidate labels or question variations), making them well-suited for cache reuse.

More broadly, the pursuit of efficient transformer inference has catalyzed a range of memory- and compute-aware optimizations. Notably, methods like FlashAttention [35] reengineer the attention mechanism itself to minimize memory access overhead and enhance throughput on modern hardware accelerators. These advancements are complementary to caching-based approaches, together contributing to a growing repertoire of scalable and latency-aware mechanisms for prompt execution.

In this work, we leverage these foundational techniques, particularly prefix caching, to support a novel dichotomic prompting formulation. This formulation entails executing a large number of structurally similar prompts efficiently, a task for which prefix caching offers substantial computational savings. Our approach situates itself within a broader trend of designing prompt strategies that are both algorithmically expressive and computationally tractable.

III. Methodology

A. LLM-to-SLM Distillation

We adopt a knowledge distillation framework [36], wherein a large teacher model supervises smaller student models to reduce inference costs while preserving annotation quality. This paradigm is particularly effective in low-resource settings, where labeled data is scarce but powerful LLMs are available [37], [38].

Our teacher model, DeepSeek-V3 [39], is a scalable mixture-of-experts decoder. It generates pseudo-labels for 24 affective dimensions using three independently sampled generations per input. Outputs are aggregated via majority vote to reduce variance and approximate a consensus annotation.

Pseudo-labels are produced using a fixed Polish-language prompt in structured JSON format:

```
Your task is to identify emotions
evoked by the given text.
1. For each emotion in the list:
{LABELS}, assess whether the text
evokes it (true/false).
2. Assess the overall tone of the text
as one or more of: positive, negative,
neutral.
The tone does not have to be
exclusive--you may assign true to
multiple options.
Return only a JSON object containing:
- all emotions with true/false values,
- three keys: positive, negative,
neutral--also with true/false values.
Text: {{text}}
```

This prompt elicits structured JSON responses with binary judgments for each of the 24 affective dimensions, along with non-exclusive polarity labels. Its format ensures consistency and facilitates downstream parsing during label aggregation and model supervision.

Figure 1 illustrates the full annotation and distillation pipeline. Raw texts are first annotated by DeepSeek-V3 using the above prompt. The aggregated labels are then verified by human annotators for quality assessment and used to fine-tune four compact models: HerBERT-Large [40], PLLuM-8B [41], Gemma3-1B [42], and an internal CLARIN-1B model.

HerBERT-Large is an encoder-only Polish language model. PLLuM-8B and CLARIN-1B are decoder-only models adapted to Polish, while Gemma3-1B is a multilingual decoder-only model supporting over 140 languages, including Polish.

B. Dichotomic Prompting Framework

Traditional multi-label classification methods, such as fixed-output architectures or structured generation, predict all labels jointly but assume a static label space, limiting their use in dynamic tasks like affective computing.

We introduce *dichotomic prompting*, which reformulates multi-label classification as K independent binary decisions, each posed as a natural-language question, where the model responds with a single token (`Yes`/`No`) that maps directly to binary labels $(1, 0)$, respectively.

Key advantages include:

- **Scalability**: Labels can be added or modified without retraining.
- **Flexibility**: Prompts can reflect subjective or evolving definitions.

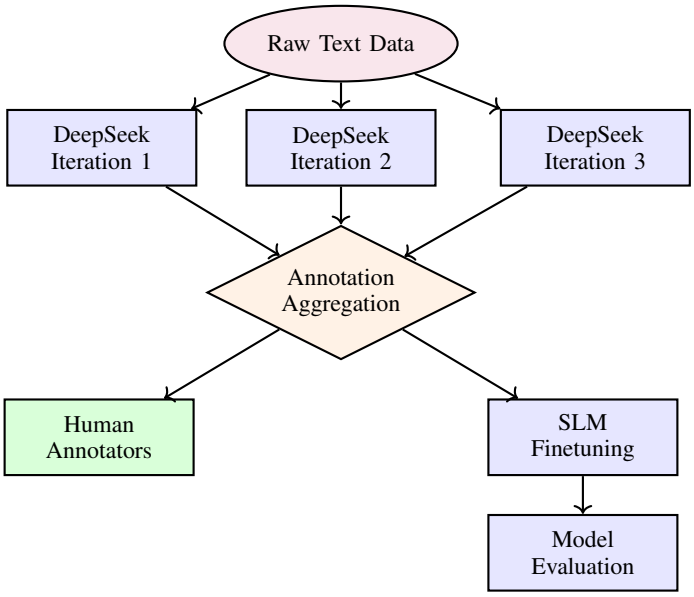

Fig. 1: Annotation and distillation pipeline. Raw texts are independently annotated via three DeepSeek-V3 passes. The outputs are aggregated (via majority vote) to form consensus pseudo-labels. These annotations are then verified by human annotators for reliability assessment and used to fine-tune small language models (SLMs), which are subsequently evaluated.

- **Efficiency**: Only relevant dimensions need to be queried.

As shown in Figure 2, dichotomic prompting differs from structured JSON prompting by issuing separate queries per label rather than generating all predictions in one pass.

Although this increases the number of forward passes, decoder-only models alleviate the cost via *prefix caching*, which reuses shared prompt segments. While we demonstrate this method for affective classification in Polish, it generalizes across languages and domains.

C. Caching and Efficient Inference

Although dichotomic prompting issues K queries per input, one per label, the overhead is greatly reduced by *prefix caching*, as supported in decoder-only models and optimized in inference frameworks such as vLLM [34]. By reusing key–value attention states for shared prompt prefixes, redundant computation is avoided across similar queries.

In our setup, each prompt contains identical instructional and input text components, differing only in the target label. Decoder-only models process these prompts in two phases: a *prefill phase* that encodes the shared prefix and a *decode phase* that generates the final token. Since each response is constrained to a single-token output (`Yes`/`No`), nearly all computation lies in the prefill phase, making caching particularly effective.

Prompt structure directly impacts cache efficiency. As illustrated in Figure 3, placing the label at the end of the prompt (Case 3) maximizes the cacheable region, as both instruction and text are reused across queries. This configuration was adopted in our main experiments.

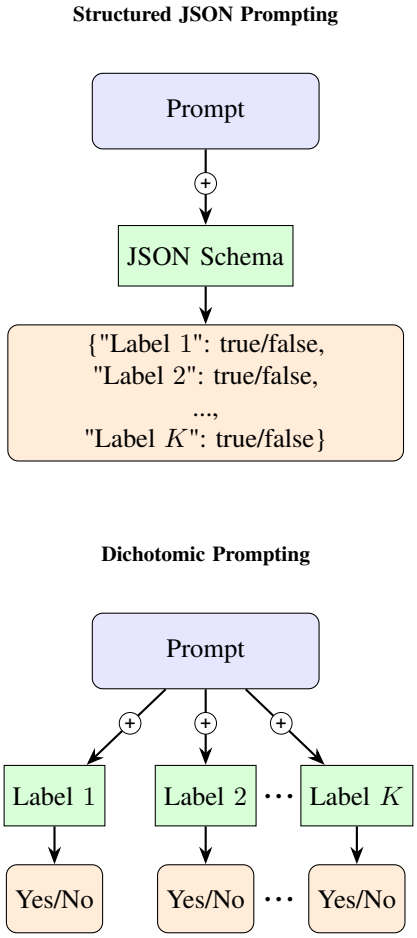

Fig. 2: Prompting strategies for multi-label classification. **Top:** Structured JSON prompting produces all label predictions in a single response by filling a predefined JSON schema. **Bottom:** Dichotomic prompting issues one binary question per label and collects individual yes/no answers. Both strategies share the same input text but differ in how labels are queried and outputs are structured.

IV. Experimental Setup

A. Dataset Composition

To evaluate our framework, we compiled a 10,000-example dataset of affective Polish-language texts, drawn from six publicly available corpora. These include `wikinews_pl` [43], `allegro_reviews` [44], `cdt` [45], `open-coursebooks-pl` [46], and a mixed-domain dataset [47], from which we extracted two subsets: `pl_twitter_sentiment` and `pl_opi_lil_2012`. These sources span multiple domains and contexts of use.

Each text was annotated with binary labels across 24 affective dimensions. These dimensions are adapted from the 26-category framework introduced by [48], who systematically derived a fine-grained set of emotions and sentiment-related

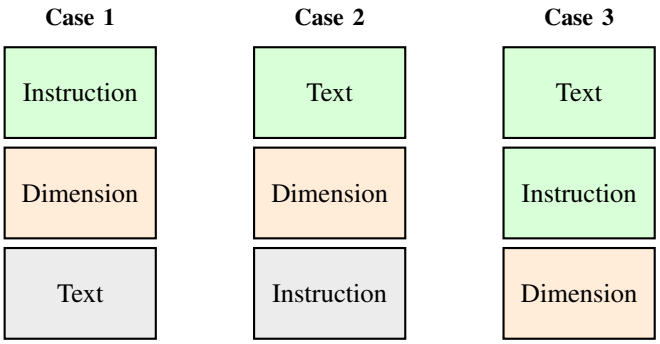

Fig. 3: Prompt structure configurations for evaluating prefix caching efficiency. Each column (Case 1–3) represents a different arrangement of prompt components: **Instruction**, **Text**, and **Target Label (Dimension)**. Green boxes denote cacheable segments reused across prompts; gray boxes indicate uncached, recomputed parts; and orange boxes highlight the position of the queried affective label. Case 3 maximizes cache utilization by placing both the instruction and input text in the shared prefix.

TABLE I: Composition of the affective text dataset used in this study, grouped by source, topic, and context of use.

Category	Type	Number of Texts
Source Corpus	`wikinews_pl`	4,633
	`allegro_reviews`	2,000
	`open-coursebooks-pl`	1,026
	`cdt`	983
	`pl_twitter_sentiment`	713
	`pl_opi_lil_2012`	645
Topic	Politics	2,000
	Sport	2,000
	Science	2,000
	Products	2,000
	Culture	2,000
Usage Context	News	4,633
	Social Media	2,341
	Review	2,000
	Coursebook	1,026

categories through large-scale questionnaires and annotation studies. This ensures that our schema remains grounded in established affective ontologies while remaining practical for large-scale Polish data.

Texts span five topical domains (*politics*, *sport*, *science*, *products*, *culture*) and four usage contexts (*news*, *social media*, *reviews*, *coursebooks*). Table I summarizes the distribution across sources, topics, and contexts.

The data is split into training, validation, and test sets with balanced label coverage. No preprocessing was applied to the text content.

B. Human and Model Annotation Agreement

In addition to fully automated LLM-based experiments, we conducted a human verification study to assess the reliability of the obtained results. Two human annotators were employed to evaluate a subset of text fragments. Their task was to determine whether any of the predefined sentiment and emotion labels were applicable to each fragment.

To ensure consistency with the simple prompt provided to the model (Section III-A), annotators were instructed to make intuitive judgments regarding the sentiment and emotional characteristics of the texts. No formal definitions or explanatory guidelines for the labels were provided. Instead, annotators were asked to identify all potential emotions or effects the text might evoke in a general reader.

In the first phase, annotators reviewed 40 text fragments accompanied by the binary label assignments produced by the LLM. They independently verified each label and corrected cases where their judgments diverged from the model's outputs. The results were encouraging: both inter-annotator agreement and annotator-LLM agreement ranged between 70% and 80%.

In a subsequent experiment, we removed the model's annotations and asked the annotators to assign label values independently (blind condition). This procedure resulted in only a slight decrease in agreement levels but increased annotation time by approximately 50%. Consequently, we adopted the initial procedure in which annotators directly verified and corrected the model's outputs.

Following this protocol, the annotators proceeded to evaluate an additional 160 text fragments.

C. Evaluation Protocol

The fine-tuned models (HerBERT-Large, CLARIN-1B, PLLuM-8B, and Gemma3-1B) are trained using binary cross-entropy loss over 24 affective dimensions. Labels are derived from the LLM-driven distillation pipeline (Section III-A) and serve as pseudo-labels generated for supervised training.

Model-specific hyperparameters are selected via 10-trial Bayesian optimization with early stopping on a held-out validation subset.

Evaluation is conducted under two regimes:

1) **In-distribution (ID):** All 24 affective labels are available in the training, validation, and test sets.
2) **Out-of-distribution (OOD):** A single label is excluded from the training and validation sets, and the model is evaluated only on that label in the test set. This leave-one-out setup is repeated across all labels.

Gemma3-1B is used as the representative model for OOD evaluation, as it is the only small decoder-only model among those tested that reliably supports both prompting formats, including structured JSON outputs.

To assess the effects of fine-tuning (e.g., catastrophic forgetting or generalization), we compare zero-shot and fine-tuned OOD performance under both prompting strategies. Prompting instructions used during evaluation are paraphrased but format-consistent with those seen during training.

Evaluation metrics include macro- and micro-averaged F1 scores. Inference latency is also measured for Gemma3-1B under both prompting styles to estimate computational efficiency.

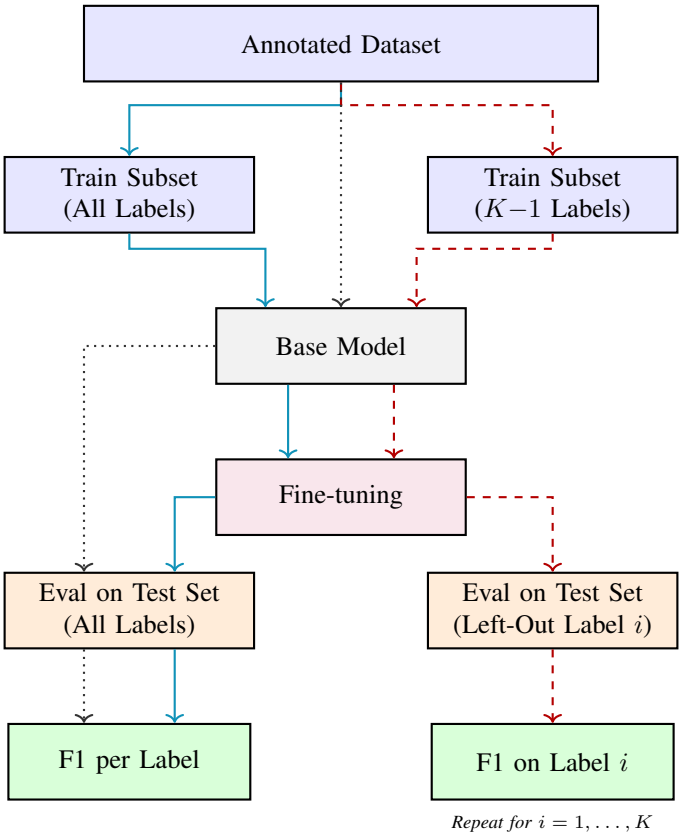

Fig. 4: Evaluation protocol for all settings.
Solid cyan path: Fine-tuning and evaluation on all labels.
Dashed red path: Leave-one-out (LOO) strategy where one label is excluded from training and evaluated separately.
Dotted gray path: Zero-shot evaluation using the base model without fine-tuning.

TABLE II: Positive Specific Agreement (PSA) scores comparing the consistency of affective annotations between human annotators and LLM-generated labels.

Comparison	PSA Score
LLM (inter-run average agreement)	0.925
Human Annotator A1 vs. A2	0.791
LLM (aggregated) vs. Annotator A1	0.818
LLM (aggregated) vs. Annotator A2	0.803

Figure 4 summarizes the evaluation setup across all configurations.

V. Results

A. Annotation Agreement Analysis

LLM-generated annotations exhibit strong alignment with human judgments. As shown in Table II, DeepSeek-V3 achieves high Positive Specific Agreement (PSA) scores against two human annotators. Inter-annotator agreement was similarly high, while intra-model consistency (inter-run) reached over 0.9.

TABLE III: Per-label in-distribution F1 scores for each model and prompting configuration. Results are grouped by model and prompting setups, where applicable. CLARIN-1B could not support structured output generation.

	HerBERT	CLARIN-1B	Gemma3-1B		PLLuM-8B	
	—	Dichotomic	JSON	Dichotomic	JSON	Dichotomic
Understandable	0.963	0.965	0.960	0.959	0.965	0.964
Trustworthy	0.962	0.958	0.929	0.971	0.967	0.968
Interesting	0.942	0.938	0.930	0.953	0.948	0.951
Positive	0.905	0.884	0.885	0.865	0.919	0.922
Negative	0.881	0.853	0.809	0.813	0.911	0.903
Neutral	0.810	0.774	0.733	0.783	0.821	0.848
Joyful	0.860	0.830	0.812	0.828	0.881	0.898
Likeable	0.832	0.834	0.797	0.812	0.851	0.867
Political	0.902	0.934	0.979	0.970	0.940	0.953
Angry	0.827	0.820	0.845	0.867	0.890	0.896
Surprising	0.490	0.564	0.194	0.306	0.757	0.655
Delightful	0.794	0.746	0.878	0.886	0.820	0.829
Inspiring	0.802	0.773	0.607	0.657	0.830	0.841
Sad	0.694	0.722	0.449	0.507	0.772	0.789
Ironic	0.647	0.615	0.619	0.633	0.812	0.805
Hurtful	0.759	0.736	0.764	0.770	0.821	0.798
Funny	0.472	0.385	0.527	0.548	0.762	0.685
Offensive	0.777	0.695	0.834	0.853	0.855	0.808
Compassionate	0.685	0.676	0.662	0.689	0.822	0.829
Vulgar	0.622	0.610	0.548	0.591	0.822	0.707
Embarrassing	0.413	0.440	0.586	0.612	0.592	0.639
Disgusting	0.469	0.585	0.626	0.635	0.666	0.529
Scary	0.232	0.316	0.516	0.521	0.688	0.645
Calm	0.000	0.070	0.200	0.208	0.343	0.427

Figure 5 shows the relationship between label prevalence and annotation agreement. We find a strong positive correlation between the frequency of positive labels and intra-model PSA ($\rho = 0.80$, $p < 0.001$), indicating that frequent categories are annotated more consistently. A similar, though weaker, correlation is observed between Annotator A1 and the model ($\rho = 0.63$, $p = 0.001$), while no significant correlation is found for Annotator A2.

B. Classification Performance

Table IV reports macro and micro F1 scores across all models and prompting strategies. Among fine-tuned models, decoder-only architectures perform on par with or slightly better than the encoder-only HerBERT. While PLLuM-8B achieves the highest scores, further analysis focuses on Gemma3-1B as a more computationally practical option.

CLARIN-1B, which lacks instruction tuning, could not reliably generate structured outputs and was evaluated only under dichotomic prompting.

Label-wise in-distribution results are shown in Table III, where dichotomic prompting performs comparably to JSON prompting across most affective dimensions and models, confirming its effectiveness despite its simplicity.

Figure 6 presents per-label F1 scores for Gemma3-1B across prompting styles and training regimes. While fine-tuning improves many dimensions, it tends to reduce generalization to unseen labels. Notably, the variation in per-label outcomes further suggests that generalization gains are not strictly determined by label frequency. In contrast, zero-shot prompting, particularly with dichotomic queries, achieves the strongest out-of-distribution performance compared to JSON prompting (Table V).

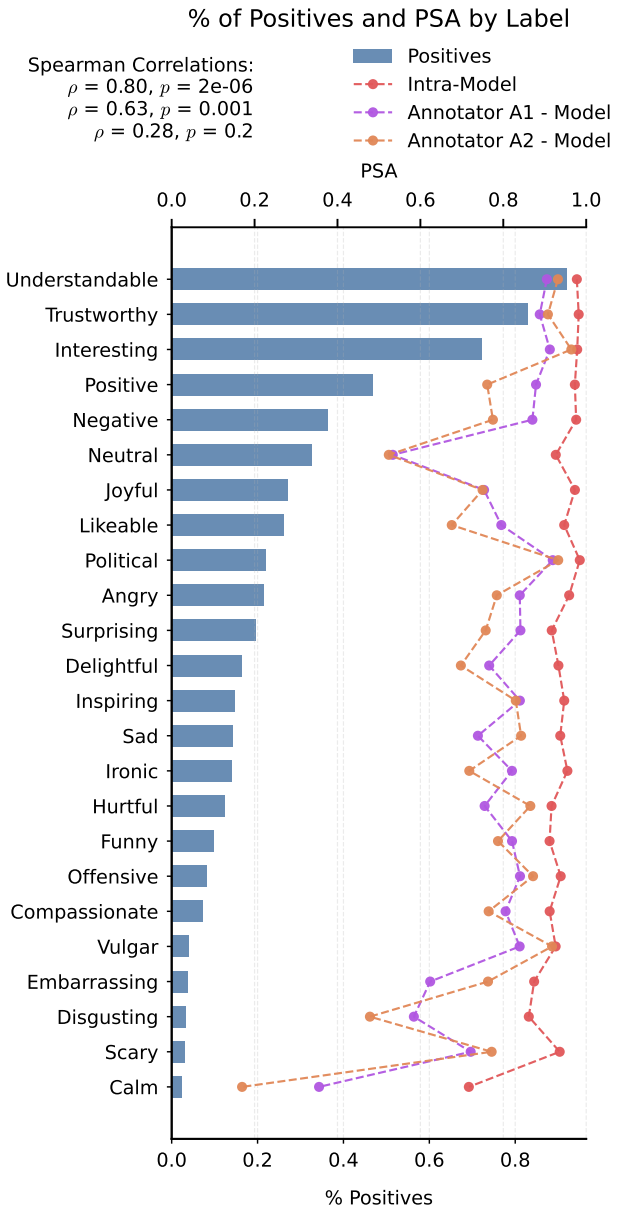

Fig. 5: Relationship between label frequency and annotation agreement. Blue bars show the percentage of positive annotations per label, while dashed lines indicate Positive Specific Agreement (PSA): intra-model consistency (red), annotator A1 vs. model (purple), and annotator A2 vs. model (orange). Reported Spearman correlations (top left) quantify the relationship between label prevalence and PSA, showing that more frequent labels tend to yield higher agreement.

These results indicate that dichotomic prompting yields more reliable zero-shot classification, overcoming common errors in structured output generation.

Finally, we evaluated the impact of prompt structure on model performance using a Wilcoxon signed-rank test over F1 scores for different prompt formats illustrated in Figure 3. The ordering of prompt components had no significant effect

TABLE IV: In-distribution macro and micro F1 scores for all evaluated models under different prompting strategies. CLARIN-1B, lacking instruction tuning, were evaluated using dichotomic prompting only.

Model	Prompting Setup	F1 (Macro)	F1 (Micro)
CLARIN-1B	Dichotomic	0.718	0.856
PLLuM-8B	JSON	0.812	0.897
	Dichotomic	0.801	0.898
Gemma3-1B	JSON	0.712	0.846
	Dichotomic	0.722	0.866
HerBERT	—	0.728	0.864

TABLE V: Out-of-distribution (OOD) classification performance of Gemma3-1B under different prompting and training regimes. F1 scores are reported for zero-shot and fine-tuned configurations, using both prompting strategies.

Training Regime	Prompting Setup	F1 (Macro)	F1 (Micro)
Zero-shot	JSON	0.250	0.300
	Dichotomic	0.363	0.445
Fine-tuned	JSON	0.299	0.340
	Dichotomic	0.286	0.178

on classification quality, confirming that dichotomic prompting is robust to structural variation.

C. Inference Efficiency

To compare the time efficiency of the two aforementioned prompting strategies, inference time was measured. The experiments were conducted on a system equipped with an NVIDIA L40 GPU featuring 45 GB of VRAM and under the CUDA version of 12.8. Inference was executed using the vLLM framework to ensure consistent model serving. Besides, vLLM supports *automatic prefix caching*. A sample of 5,000 texts was used for evaluation, with text lengths ranging from 300 to 8,000 tokens. All the measurements were collected for the Gemma3-1B model.

The efficiency in the dichotomic prompting scenario varies significantly depending on the position of the affective dimension within the prompt, as presented in Figure 7. This effect is primarily driven by the functioning of the prefix caching mechanism. The longer the shared prefix, the greater the portion of the prompt that can be cached, which in turn reduces the computational cost for subsequent prompts that share the same text but differ in the evaluated dimension.

The difference is particularly pronounced between scenarios where the text is not cached (Case 1) and those where it is (Case 2 and Case 3). Since the text typically constitutes the longest part of the prompt, placing the dimension before the text prevents caching of the majority of the input. Comparing Case 2 with Case 3 further confirms that positioning the dimension as close to the end of the prompt as possible yields faster evaluation, although the difference here is less substantial. This pattern remains consistent across different batch sizes. As our earlier experiments demonstrated that

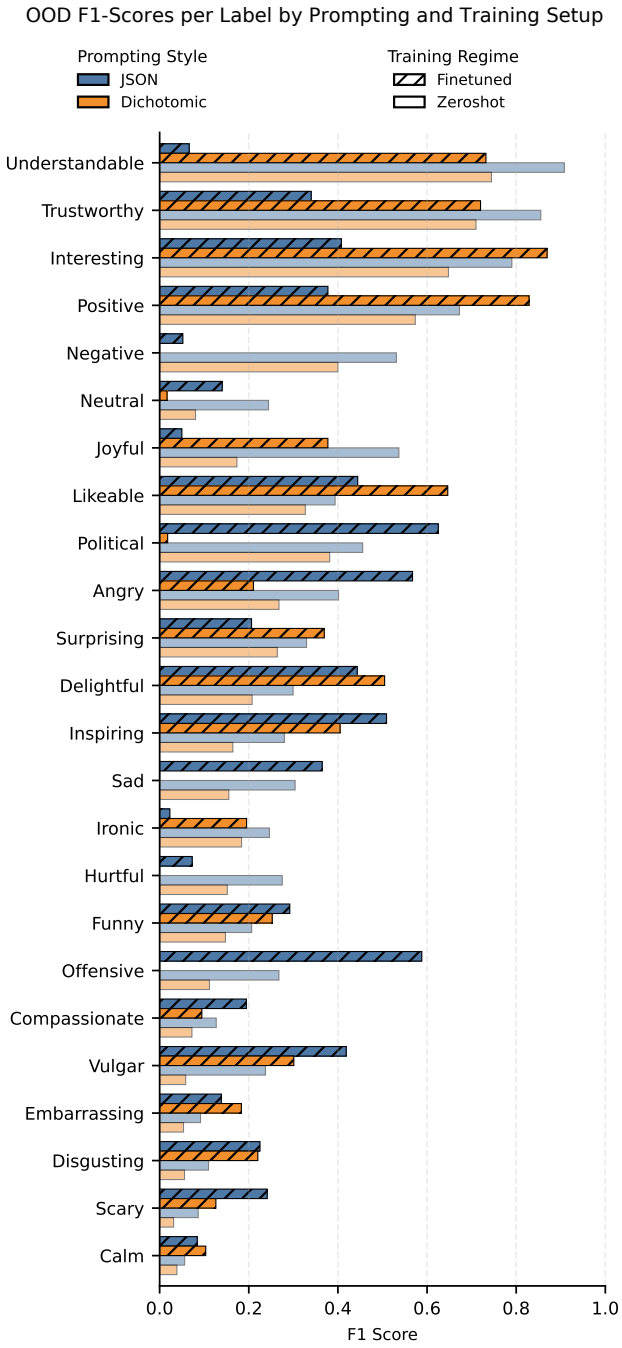

Fig. 6: Per-label F1 scores on the Out-of-Distribution (OOD) setup using Gemma3-1B under different prompting styles and training regimes.

evaluation quality remains comparable across all three cases, we recommend adopting Case 3 as the preferred approach.

The comparative effectiveness of Dichotomic versus JSON prompting varies with the length of the input texts, as illustrated in Figure 8. Dichotomic prompting is particularly efficient for shorter texts, but its advantage over JSON prompting diminishes as the input texts become longer.

This suggests that Dichotomic prompting is advantageous

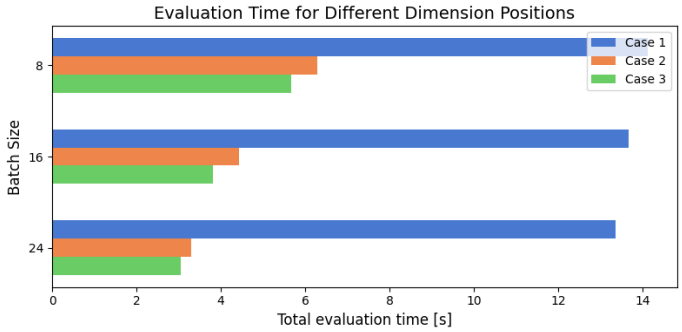

Fig. 7: Comparison of evaluation times based on the placement of the affective dimension in the prompt using Gemma3-1B. Reported values correspond to the total evaluation time for a sample of 1,000 texts, each 300 tokens in length.

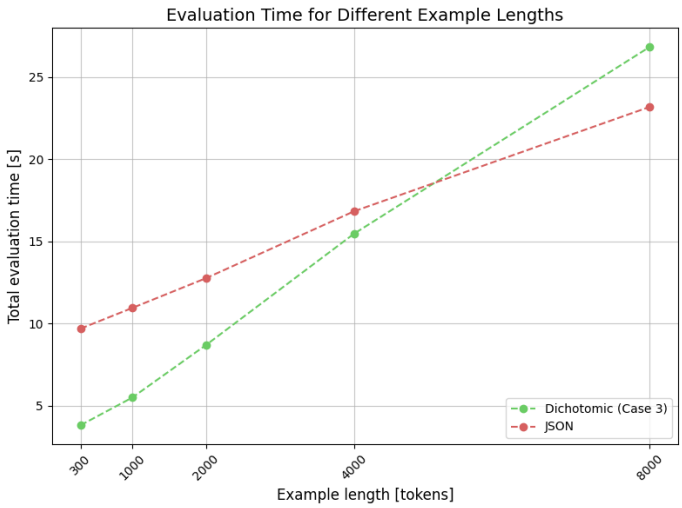

Fig. 8: Comparison of evaluation times between Dichotomic and JSON prompting strategies using Gemma3-1B. Reported values correspond to the total evaluation time for a sample of 1,000 texts at each of the five examined text lengths.

for short texts, such as those typically used in affective evaluations, and in scenarios with limited computational resources, where batch sizes are necessarily small. We hypothesize that the diminishing advantage of Dichotomic prompting with longer texts and larger batch sizes is due to the limited capacity of the cache: longer texts cannot all be stored, and larger batches contain more texts, making it harder to keep them in the cache and thereby reducing the relative efficiency of Dichotomic prompting. A more detailed investigation of this phenomenon is left for future work.

VI. Discussion

Our results highlight several key insights into dichotomic prompting for multi-label classification with decoder-only language models. Fine-tuned small models, especially PLLuM-8B, perform well on in-distribution data, but the generalization to out-of-distribution labels is limited. In contrast, zero-shot

prompting (prior to fine-tuning) shows greater robustness to label exclusion. This illustrates the trade-off between in-domain optimization and generalization capacity.

Dichotomic prompting matches structured JSON prompting across models in both overall and per-label performance, despite its simpler format. This holds even under prompt structure variations, with no significant accuracy loss.

Efficiency-wise, dichotomic prompting benefits from prefix caching, enabling faster inference on short texts by reusing shared prompt segments across label queries. While structured prompting processes all labels in one pass, it lacks caching opportunities. For short inputs, dichotomic prompting is faster; however, its advantage wanes with longer texts. These trade-offs point to hybrid strategies, adapting the prompting format to input length or label set, as promising paths for deployment.

VII. Limitations

Despite strong results, several limitations remain.

First, the annotation process depends entirely on a single teacher model, which, despite strong agreement with human annotations, may introduce biases or inconsistencies that propagate during training. Moreover, our human verification covered only a relatively small subset (200 texts), which may not be sufficient to fully capture potential systematic biases.

Second, the speedups from prefix caching are hardware- and system-dependent. Our setup favored dichotomic prompting for short inputs, but this advantage diminishes for longer texts due to cache capacity limits, and results may vary across platforms, models, or batch sizes.

Lastly, this work focuses on affective classification in Polish. While the method is language-agnostic in principle, its generalization to other domains, languages, or annotation schemas remains untested. The conclusions should thus be interpreted within this scope.

VIII. Future Work

Future work should explore scaling dichotomic prompting to tasks with larger label spaces. While this study focused on 24 affective dimensions, many applications, like topic classification or clinical coding, involve hundreds of labels. Understanding how performance and caching efficiency scale remains an open question.

Broader evaluation across languages and domains is also needed. Cross-lingual and cross-domain testing would provide insights into the method's adaptability and robustness.

Hybrid prompting strategies also merit investigation. Dynamically choosing between structured and dichotomic setups based on input length, label count, or compute resources could optimize both speed and flexibility.

IX. Conclusion

We presented a modular framework for multi-label classification using dichotomic prompting with decoder-only language models. By framing the task as a sequence of binary decisions and exploiting prefix caching, the approach balances accuracy, efficiency, and adaptability.

Experiments show that small models fine-tuned on LLM-generated labels can match or exceed zero-shot performance on in-domain data. Zero-shot prompting, meanwhile, generalizes better to unseen labels. Dichotomic prompting performs comparably to structured JSON prompting across all settings, while offering added benefits for short inputs.

Overall, the framework provides a scalable and efficient solution for multi-label classification, well-suited to dynamic label spaces and constrained compute environments.

Acknowledgments

This work was financed by (1) the European Regional Development Fund as part of the 2021–2027 European Funds for a Modern Economy (FENG) programme, project no. FENG.02.04-IP.040004/24: CLARIN – Common Language Resources and Technology Infrastructure; (2) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, project no. POIR.04.02.00-00C002/19: CLARIN – Common Language Resources and Technology Infrastructure; (3) Digital Research Infrastructure for the Arts and Humanities DARIAH-PL (POIR.04.02.00-00-D006/20, KPOD.01.18-IW.03-0013/23); (4) CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Minister of Science under the programme: "Support for the participation of Polish scientific teams in international research infrastructure projects", agreement number 2024/WK/01; (5) Polish Minister of Digital Affairs under a special purpose subsidy No. 1/WII/DBI/2025: HIVE AI: Development and Pilot Deployment of Large Language Models in the Polish Public Administration; (6) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology.

AI-based tools, including ChatGPT, Grammarly Premium, and Writeful, were used exclusively to support linguistic clarity and improve the readability of the manuscript.

References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, p. e12189, 2020.

[3] M. Kmak, K. Chmurzyński, K. Matejuk, P. Kotzbach, and J. Kocoń, "Predicting stock prices with chatgpt-annotated reddit sentiment: Hype or reality?" in *International Conference on Computational Science*. Springer, 2025, pp. 307–322.

[4] A. Ngo and J. Kocoń, "Integrating personalized and contextual information in fine-grained emotion recognition in text: A multi-source fusion approach with explainability," *Information Fusion*, vol. 118, p. 102966, 2025.

[5] T. Ferdinan and J. Kocoń, "Fortifying nlp models against poisoning attacks: The power of personalized prediction architectures," *Information Fusion*, vol. 114, p. 102692, 2025.

[6] Ł. Radliński, M. Guściora, and J. Kocoń, "Backtranslation and paraphrasing in the llm era? comparing data augmentation methods for emotion classification," in *International Conference on Computational Science*. Springer, 2025, pp. 3–17.

[7] A. Karlińska, P. Miłkowski, P. Czwordon-Lis, B. Koptyra, and J. Kocoń, "Comprehensive sentiment analysis of polish book reviews using large and small language models," in *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2024, pp. 453–462.

[8] B. Koptyra, M. Oleksy, E. Dzięcioł, and J. Kocoń, "Small language models for emotion recognition in polish stock market investor opinions," in *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2024, pp. 463–470.

[9] J. Kocoń, "Deep emotions across languages: A novel approach for sentiment propagation in multilingual wordnets," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2023, pp. 744–749.

[10] A. Ngo, A. Candri, T. Ferdinan, J. Kocoń, and W. Korczynski, "Studemo: A non-aggregated review dataset for personalized emotion recognition," in *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022, pp. 46–55.

[11] P. Miłkowski, S. Saganowski, M. Gruza, P. Kazienko, M. Piasecki, and J. Kocoń, "Multitask personalized recognition of emotions evoked by textual content," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2022, pp. 347–352.

[12] J. Kocoń, P. Miłkowski, M. Wierzba, B. Konat, K. Klessa, A. Janz, M. Riegel, K. Juszczyk, D. Grimling, A. Marchewka *et al.*, "Multilingual and language-agnostic recognition of emotions, valence and arousal in large-scale multi-domain text reviews," in *Language and Technology Conference*. Springer, 2019, pp. 214–231.

[13] J. Kocoń, M. Zaśko-Zielińska, and P. Miłkowski, "Multi-level analysis and recognition of the text sentiment on the example of consumer opinions," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 2019, pp. 559–567.

[14] K. Gawron, M. Pogoda, N. Ropiak, M. Swędrowski, and J. Kocoń, "Deep neural language-agnostic multi-task text classifier," in *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2021, pp. 136–142.

[15] W. Korczyński and J. Kocoń, "Compression methods for transformers in multidomain sentiment analysis," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 419–426.

[16] J. Szołomicka and J. Kocon, "Multiaspectemo: Multilingual and language-agnostic aspect-based sentiment analysis," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 443–450.

[17] J. Kocoń and A. Janz, "Propagation of emotions, arousal and polarity in wordnet using heterogeneous structured synset embeddings," in *Proc. of the 10th Global Wordnet Conference*, 2019, pp. 336–341.

[18] J. Kocoń, J. Baran, K. Kanclerz, M. Kajstura, and P. Kazienko, "Differential dataset cartography: Explainable artificial intelligence in comparative personalized sentiment analysis," in *International Conference on Computational Science*. Springer, 2023, pp. 148–162.

[19] J. Kocoń, A. Janz, and M. Piasecki, "Context-sensitive sentiment propagation in wordnet," in *Proceedings of the 9th global wordnet conference*, 2018, pp. 329–334.

[20] J. Kocoń, J. Radom, E. Kaczmarz-Wawryk, K. Wabnic, A. Zajączkowska, and M. Zaśko-Zielińska, "Aspectemo: multi-domain corpus of consumer reviews for aspect-based sentiment analysis," in *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2021, pp. 166–173.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[22] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=gEZrGCozdqR

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[24] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[25] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, pp. 64–74, 2008.

[26] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," *Data mining and knowledge discovery handbook*, pp. 667–685, 2010.

[27] X.-R. Yu, D.-B. Wang, and M.-L. Zhang, "Partial label learning with emerging new labels," *Machine Learning*, vol. 113, no. 4, pp. 1549–1565, 2024.

[28] A. N. Tarekegn, M. Ullah, and F. A. Cheikh, "Deep learning for multi-label learning: A comprehensive survey," *arXiv preprint arXiv:2401.16549*, 2024.

[29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[30] Y. Kementchedjhieva and I. Chalkidis, "An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 5828–5843.

[31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM computing surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[32] L. Galke, A. Diera, B. X. Lin, B. Khera, T. Meuser, T. Singhal, F. Karl, and A. Scherp, "Are we really making much progress in text classification? a comparative review," *arXiv preprint arXiv:2204.03954*, 2022.

[33] Y. Sun, L. Dong, Y. Zhu, S. Huang, W. Wang, S. Ma, Q. Zhang, J. Wang, and F. Wei, "You only cache once: Decoder-decoder architectures for language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 7339–7361, 2024.

[34] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th symposium on operating systems principles*, 2023, pp. 611–626.

[35] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in neural information processing systems*, vol. 35, pp. 16 344–16 359, 2022.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[37] A. M. Mansourian, R. Ahmadi, M. Ghafouri, A. M. Babaei, E. B. Golezani, Z. Y. Ghamchi, V. Ramezanian, A. Taherian, K. Dinashi, A. Miri *et al.*, "A comprehensive survey on knowledge distillation," *arXiv preprint arXiv:2503.12067*, 2025.

[38] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou, "A survey on knowledge distillation of large language models," *arXiv preprint arXiv:2402.13116*, 2024.

[39] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[40] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik, "Herbert: Efficiently pretrained transformer-based language model for polish," in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 2021, pp. 1–10.

[41] PLLuM Consortium, "Pllum: A family of polish large language models," https://huggingface.co/CYFRAGOVPL, 2025.

[42] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[43] R. Poświata, "wikinews-pl," https://huggingface.co/datasets/rafalposwiata/wikinews-pl, 2021.

[44] P. Rybak, R. Mroczkowski, J. Tracz, and I. Gawlik, "Klej: Comprehensive benchmark for polish language understanding," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1191–1201.

[45] M. Ptaszyński, A. Pieciukiewicz, and P. Dybała, "Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter," in *Proceedings of the PolEval 2019 Workshop*. Warsaw, Poland: Institute of Computer Science, Polish Academy of Sciences, 2019.

[46] R. Poświata, "open-coursebooks-pl," https://huggingface.co/datasets/rafalposwiata/open-coursebooks-pl, 2021.

[47] Brand24 AI Lab, "mms: Polish online media sentiment corpus," https://huggingface.co/datasets/Brand24-AI/mms, 2023.

[48] W. Mieleszczenko-Kowszewicz, K. Kanclerz, J. Bielaniewicz, M. Oleksy, M. Gruza, S. Wozniak, E. Dziecіol, P. Kazienko, and J. Kocon, "Capturing human perspectives in nlp: Questionnaires, annotations, and biases." in *NLPerspectives@ ECAI*, 2023.