What properties of reasoning supervision are associated with improved downstream model quality?

Mikołaj Langner[0009−0007−9531−5329], Dzmitry Pihulski[0009−0002−5434−4696], Jan Eliasz[0009−0007−0851−1816], Michał Rajkowski[0009−0003−2983−4324], Przemysław Kazienko[0000−0001−5868−356X], Maciej Piasecki[0000−0003−1503−0993], Jan Kocoń[0000−0002−7665−6896], and Teddy Ferdinan[0000−0003−3701−3502]

Wroclaw Tech, 50-370 Wrocław, Poland
{mikolaj.langner, dzmitry.pihulski, jan.eliasz, michal.rajkowski, kazienko, maciej.piasecki, jan.kocon, teddy.ferdinan}@pwr.edu.pl

Abstract. Validating training data for reasoning models typically requires expensive trial-and-error fine-tuning cycles. In this work, we investigate whether the utility of a reasoning dataset can be reliably predicted prior to training using intrinsic data metrics. We propose a suite of quantitative measures and evaluate their predictive power by fine-tuning 8B and 11B models on semantically distinct variants of a Polish reasoning dataset. Our analysis reveals that these intrinsic metrics demonstrate strong and significant correlations with downstream model performance. Crucially, we find that the predictors of utility are scale-dependent: smaller models rely on alignment-focused metrics to ensure precision, whereas larger models benefit from high redundancy, utilizing verbose traces to solve complex tasks. These findings establish a scale-aware framework for validating reasoning data, enabling practitioners to select effective training sets without the need for exhaustive empirical testing.

Keywords: Large Language Models · Reasoning · Data Quality · Dataset Evaluation

1 Introduction

Explicit reasoning strategies [36] and specialized models [26,12] have transformed the capabilities of Large Language Models (LLMs). Consequently, fine-tuning on datasets enriched with reasoning traces has become the standard paradigm for imbuing these models with such skills. However, while the importance of high-quality data is universally acknowledged, the definition of quality for reasoning traces remains ambiguous.

Currently, validating a reasoning dataset is an inefficient process that relies on post-hoc evaluation: researchers must fine-tune a model to discover if their data is effective. This *training-as-validation* approach is computationally prohibitive and unscalable. To democratize the development of robust reasoning models,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

the community requires objective and computable metrics that can validate the utility of training data **before** the expensive fine-tuning process begins.

In this paper, we address this gap by establishing a link between intrinsic data characteristics and downstream model performance. We leverage a controlled set of Polish reasoning variants from our previous work [29] and the corresponding fine-tuned 8B and 11B models. By subjecting these known reasoning variants to a rigorous set of quantitative measurements, from linguistic complexity to semantic alignment, we determine which metrics serve as reliable predictors of a model's final reasoning ability.

Our analysis is guided by the following research questions:

RQ1: Is it feasible to validate the utility prior to fine-tuning?

RQ2: Which specific quantitative measures provide the most meaningful signal for validating training data quality?

The contributions of this work are as follows: (1) a systematic evaluation of validation metrics for reasoning datasets, distinguishing between superficial statistics and deep semantic indicators; (2) a correlation analysis linking pre-training data scores with downstream performance; and, (3) a scale-aware framework for selecting reasoning data, allowing researchers to estimate model performance without incurring the full cost of training.

1.1 Related Work

The precise utility of generated *reasoning traces* remains a subject of active debate. Shojaee et al. [32] argue that reasoning-augmented models often exhibit *illusionary* improvements, failing catastrophically on complex tasks while *overthinking* simple ones. Although Lawsen et al. [20] challenged these findings based on methodological discrepancies, the consensus remains that reasoning traces are not a guaranteed panacea. Furthermore, studies indicate that LLM reasoning often diverges from genuine logical inference [7,5,17,38,3,10,6,29], with models frequently omitting premises or generating hallucinated reasoning steps that do not correlate with the accuracy of the final answer.

Recent work has attempted to isolate specific attributes of reasoning data that drive performance, particularly sequence length. Jin et al. [16] posit that extending the length of reasoning, regardless of quality, can boost performance. In contrast, Wu et al. [39] demonstrate an inverted U-shaped relationship, suggesting that excessive length introduces error accumulation.

Collectively, these conflicting findings suggest that neither *length* nor *presence of reasoning* alone are sufficient proxies for the utility of training data. Although prior work largely evaluates reasoning quality by analyzing model outputs, there is insufficient research on validation methods that evaluate reasoning data before committing computational resources to fine-tuning. Our work addresses this gap by correlating intrinsic data metrics with downstream performance established in our prior experiments.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

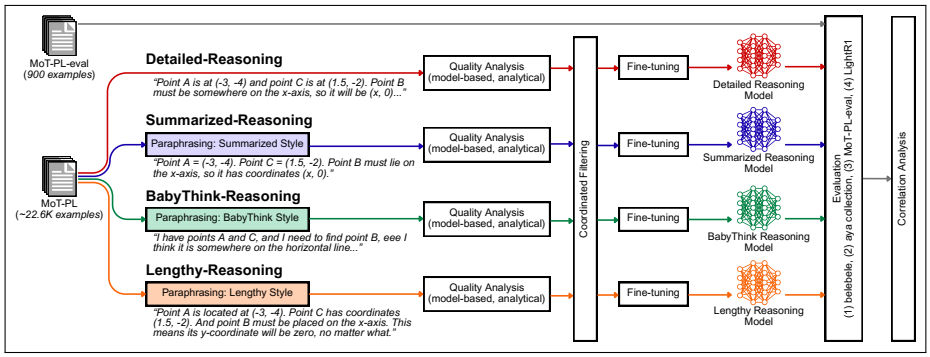

Fig. 1: We translated a subset of Mixture-of-Thoughts [13,23,27,2] into Polish, and split it into a training (MoT-PL) and evaluation set (MoT-PL-eval). Three additional variants of MoT-PL were created by paraphrasing only the reasoning part of each example: the *Summarized* style made the reasoning much more concise, the *BabyThink* style greatly simplified the reasoning, and the *Lengthy* style prolonged the reasoning. Afterwards, we fine-tuned PLLuM-8B-instruct and Bielik-11B-v2.6-Instruct on these datasets separately and evaluated them.

2 Experimental Setup

2.1 Datasets

To rigorously evaluate the efficacy of pre-training validation metrics, four distinct reasoning datasets derived from the **Polish Mixture-of-Thoughts (MoT-PL)** were used. The original MoT-PL dataset was created by sampling approximately 32,000 examples from the English Mixture-of-Thoughts collection [13] and translating them into Polish using DeepSeek-V3 [9]. After filtering for errors and context length, the final dataset contained 22,571 examples spanning three domains: Mathematics (28%), Programming (17%), and Science (55%). To ensure the generated traces exhibited natural, human-like fluency rather than rigid machine translation artifacts, a randomly sampled subset of the DeepSeek-V3 outputs was manually verified by native Polish speakers.

From this foundational dataset, we generated four semantically distinct variants to serve as our controlled variables (see Figure 1). These datasets, *Detailed*, *Summarized*, *BabyThink*, and *Lengthy*, share identical user prompts and final answers but differ significantly in the style, length, and semantic density of their reasoning traces. The general statistics of the dataset variants are shown in Table 1. The variants were generated by automatic paraphrasing using DeepSeek-V3, resulting in the following profiles:

- **Detailed:** The unmodified MoT-PL dataset, representing high-quality standard reasoning. The traces mimic the depth of the original English Mixture-of-Thoughts, serving as our control for *standard* reasoning density.
- **Summarized:** A concise variant in which reasoning traces were compressed to retain essential logic while stripping stylistic fluff. This dataset tests the hypothesis that a higher information density correlates with efficiency.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

- **BabyThink:** A variant paraphrased into "childlike" language. Rather than merely reducing statistical readability, the prompt intentionally obfuscates specific details and calculations with vague filler. The original train of thought and structure are strictly preserved to avoid injecting artificial hallucinations or new reasoning fallacies.
- **Lengthy:** An artificially prolonged variant designed to be approximately twice as long as the *Detailed* version. It preserves the original logic but introduces verbosity, allowing us to test if metrics favoring longer chains are misleading.

Table 1: Statistical profile of the dataset variants used for metric validation. All variants share identical question/answer pairs; variations occur strictly within the reasoning trace. The first value of a token count comes from using the PLLuM-8B-instruct tokenizer, while the second value comes from using the Bielik-11B-v2.6-Instruct tokenizer.

Metric (Avg.)	**Detailed**	**Summarized**	**BabyThink**	**Lengthy**
Reasoning Tokens	2,613~2,918	399~452	3,729~4,000	9,930~10,902
Reasoning Chars	6,334	965	8,991	17,369
Total Seq. Tokens	3,288~3,685	1,073~1,219	4,404~4,767	10,604~11,669
Total Seq. Chars	8,090	2,721	10,747	25,938

All examples exceeding the context window limit (32k tokens) were filtered out prior to statistical analysis and training to ensure a consistent evaluation across all variants.

2.2 Target Models

To establish a robust performance baseline across different architectures, we utilized two state-of-the-art Polish-centric LLMs as a backbone for our experiments:
- **PLLuM-8B-instruct** [18,28,19]: A derivative of Llama-3.1-8B [11], adapted via continual pre-training and instruction tuning on a massive Polish corpus;
- **Bielik-11B-v2.6-Instruct** [25]: Built upon Mistral 7B v0.2 [15], similarly enhanced with Polish-specific pre-training and fine-tuning.

Since none of the model possesses native reasoning capabilities, we adapted them by introducing special tokens `<think>` and `</think>` and expanding their embedding layers accordingly. The models were fine-tuned separately on four dataset variants (Section 2.1), resulting in a diverse set of checkpoints with varying reasoning behaviors. The technical specifications are detailed in Appendix 8.

2.3 Downstream Performance Benchmarks

To measure the utility of the training data variants, we evaluated the fine-tuned models on a comprehensive suite of benchmarks. These evaluation scores serve as **ground truth labels** against which we correlate our pre-training data metrics.

We selected four diverse benchmarks to capture different aspects of reasoning and language understanding:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

– **MoT-PL-eval**: The held-out test split of our **MoT-PL** dataset (see Section 2.1), serving as the primary metric in-domain for Polish reasoning.
– **Belebele** [1]: A challenging multilingual reading comprehension benchmark testing the models' ability to extract information from complex passages.
– **Aya Collection** [33]: A broad instruction-following suite covering summarization, classification, and QA, used to verify general capability retention.
– **LightR1** [37]: An English-language benchmark for high-difficulty logical tasks, included to assess the transfer of cross-lingual reasoning.

2.4 Evaluation Protocol

To obtain the ground-truth performance scores needed for our correlation analysis, we evaluated all fine-tuned models on the four benchmarks described above. For each dataset, we sampled a stratified test set of 900 examples to ensure balanced coverage of reasoning lengths and task types. We report each model performance using two primary metrics: Absolute Accuracy and Relative Percentage Change compared to the base model, to isolate the specific impact of training data.

Given the scale of evaluation, we adopted the *LLM-as-a-judge* paradigm. We used `DeepSeek-R1-0528` [12] as an oracle judge. The judge was strictly prompted to assess the correctness of the final answer (ignoring intermediate reasoning steps) against the ground truth. This binary decision process was applied across all benchmarks.

To ensure the reliability of these generated scores, we conducted a manual audit on a subset of 100 random samples from the **MoT-PL-eval** dataset. A human expert annotated these samples blindly (without seeing the model's judgment). The agreement rate between the human annotator and `DeepSeek-R1-0528` was **95%**, with a Cohen's Kappa score of **0.886**. This strong alignment confirms that our automated ground-truth labels are a reliable proxy for human evaluation.

During evaluation, the judge was provided with the query, reference answer, and model prediction, and instructed to output a binary decision in a constrained JSON format. The exact prompt templates for all benchmarks are available in our public repository[1].

3 Methodology

To systematically evaluate the utility of reasoning data prior to training, we propose a multi-dimensional validation framework. We categorize our metrics into two distinct groups: **Model-based Metrics** and **Analytical Metrics**. With Model-based Metrics, we aim to assess the logical integrity of the reasoning trace. We adopted the FVCU (Factuality, Validity, Coherence, Utility) taxonomy proposed by [21] for these metrics. Meanwhile, we designed our Analytical Metrics to measure statistical and structural properties of the text.

[1] https://github.com/DzmitryPihulski/prompts

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

3.1 Model-based Metrics

To assess the intrinsic quality of the reasoning steps beyond binary correctness, we implement an automated evaluation pipeline based on the **FVCU** taxonomy (Factuality, Validity, Coherence, Utility) [21]. This approach verifies whether the reasoning process itself is sound at the atomic level.

We utilize a two-stage pipeline consisting of an *Atomizer* and a *Judge*, both powered by **Qwen3-235B-A22B-Instruct-2507-FP8** [30].

1. **Atomizer:** Decomposes raw reasoning traces into atomic steps using a strict verbatim extraction strategy. This preserves the original density and style of the text, aligning with process supervision standards [22].
2. **Judge:** Evaluates each step, one-by-one, against the FVCU taxonomy.

Metric Definitions

- **Factuality (F):** Assesses the consistency with premises and external truths using the *Principal Knowledge Grounding* method [14], ensuring that steps are supported by explicit problem statements rather than hallucinated constraints.
- **Validity (V):** Evaluates the mathematical and inferential correctness of the derivation. It distinguishes between calculation errors and logical fallacies.
- **Coherence (C):** Checks if the step logically follows the preceding one without gaps, satisfying the Markov property of the chain [35].
- **Utility (U):** Measures whether the step contributes effective progress towards the solution, distinguishing constructive decomposition from "reasoning loops".

3.2 Analytical Metrics

To complement the computationally expensive FVCU, we compute scalable structural metrics across the full training dataset:

- **Semantic Alignment:** Cosine similarity between query and reasoning trace embeddings (using `mmlw-roberta-large` [8]), serving as a proxy for instruction adherence [40].
- **Semantic Flow:** Average cosine similarity between consecutive sentences, quantifying narrative smoothness, and transitional logic [40].
- **Redundancy Ratio:** Information density calculated as $\left(1 - \frac{\text{len}_{compressed}}{\text{len}_{original}}\right)$, using `zlib` compression. Higher values indicate repetitive patterns or verbosity [31,4].
- **Syntactic Depth:** Average maximum depth of dependency trees (computed via `spacy` library), indicating linguistic complexity and cognitive load [40].
- **Symbolic Fraction:** Ratio of non-alphanumeric characters to total text, capturing the density of mathematical or code-like notation [31,4].
- **Perplexity:** Exponentiated average negative log-likelihood per token taken from `Qwen3-4B` [30], measuring the text's conformity to general knowledge [31,4].

A core motivation of our framework is replacing costly trial-and-error fine-tuning with efficient pre-training validation. While brute-force empirical validation requires heavy forward and backward passes across all candidate datasets, incurring massive computational debt, our analytical pipeline bypasses gradient

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

updates entirely. By relying strictly on lightweight processing and single-pass embedding extraction, we reduce the validation footprint from dozens of multi-GPU hours to negligible compute time.

4 Results and Analysis

In this section, we present the empirical findings of our validation study. We begin by analyzing the intrinsic quality of the datasets using our framework: first, the model-based evaluation on a 1,000 subsample and second, the analytical profiling of the full training corpora. Finally, we report the performance of the model in downstream tasks and correlate these metrics to identify the most reliable predictors of success.

4.1 Model-based Metrics

Due to the prohibitive computational cost of model-based judging, FVCU metrics were evaluated on a single subsample of 1,000 examples per variant. To mitigate the variance inherent in single-batch evaluation, we employed rigorous stratified sampling, ensuring the subset accurately preserves the domain and complexity distribution of the full dataset. While the lack of multiple independent batches precludes formal variance calculations, this stratified design yields a highly representative estimate. Consequently, we frame these FVCU scores not as absolute statistical bounds, but as robust directional indicators of the reasoning trade-offs between our dataset variants. Table 2 presents these results.

Table 2: Model-based metrics evaluation on 1,000 **MoT-PL** subsamples.

Dataset	Factuality	Validity	Coherence	Utility
BabyThink	78.0	65.8	91.4	54.4
Detailed	95.0	94.3	97.5	84.5
Lengthy	94.8	95.2	98.5	87.0
Summarized	92.2	87.3	96.7	90.5

The **Summarized** variant maximizes Utility (90.5%) but at the expense of Validity (87.3%). This expected drop in Validity occurs because our LLM judge strictly evaluates explicit step-by-step derivation, penalizing the intentional omission of intermediate steps as logical gaps even when the final conclusion remains factual and highly useful. In contrast, **Lengthy** achieves the highest Validity (95.2%) and Coherence (98.5%), indicating that granular, explicit derivations are essential for stabilizing the reasoning process. Finally, the baseline **BabyThink** demonstrates that high Coherence (91.4%) is insufficient for reasoning quality — its low Validity (65.8%) confirms that the model can generate linguistically smooth but factually ungrounded chains.

These findings highlight a critical trade-off in reasoning data curation: while stripping intermediate steps (**Summarized**) increases immediate task utility, it degrades the rigorous logical grounding required for out-of-distribution generalization. Conversely, verbosity (**Lengthy**) acts as a safeguard against hallucination

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

by enforcing strict state-tracking, which is essential for complex reasoning but requires sufficient model capacity to process.

4.2 Analytical Metrics

We extended our analysis to the entire training dataset using computationally efficient metrics. Table 3 summarizes the profiles of each variant.

Table 3: Analytical metrics calculated on the full training **MoT-PL** datasets.

Dataset	Syntactic Depth	Sem. Flow	Sem. Align.	Perplexity	Redundancy Ratio	Symbolic Fraction
BabyThink	3.22	0.861	0.916	2.42	0.622	0.098
Detailed	4.41	0.868	0.951	1.41	0.623	0.131
Lengthy	4.61	0.856	0.941	1.76	0.629	0.127
Summarized	4.92	0.881	0.950	1.38	0.441	0.201

The *Summarized* dataset emerges as the most information-dense, exhibiting the highest **Symbolic Fraction** (0.201) and **Syntactic Depth** (4.92) while maintaining the lowest **Redundancy Ratio** (0.441). In contrast, the *Lengthy* and *Detailed* variants share nearly identical redundancy scores ($\sim$0.62), suggesting that the *Lengthy* variant scales volume without altering the fundamental compression rate of the text. Notably, *BabyThink* variant, despite its simplified vocabulary, yields the highest **Perplexity** (2.42) and the lowest **Semantic Alignment** (0.916).

These structural differences imply that reasoning quality is not merely a function of length but of information pacing. The high Perplexity and low Semantic Alignment of the *BabyThink* variant suggest that artificially simplifying vocabulary disrupts the natural language distribution the model expects, paradoxically making the reasoning harder to learn from despite its simpler syntax.

4.3 Downstream Model Performance

We evaluate the fine-tuned models across four benchmarks to establish the ground truth for our correlation analysis. We present the results in three stages: Absolute Accuracy, Relative Performance Change, and finally domain-specific breakdown.

Table 4 presents the absolute accuracy. Consistent with the difference in model size, Bielik-11B significantly outperforms PLLuM-8B. For **PLLuM-8B**, the *Detailed* variant achieves the highest average performance (0.513), showing particular strength in Polish reasoning tasks on MoT-PL-eval (0.374). For **Bielik-11B**, the *Lengthy* variant emerges as the superior specialist in reasoning overall, achieving the highest absolute scores on both MoT-PL-eval (0.701) and LightR1 (0.599).

Table 5 reports the Relative Percentage Change to normalize for the base model capabilities. The **Detailed** model proved to be the safest strategy, delivering consistent gains for PLLuM-8B on most benchmarks. Including general NLP benchmarks in the average demonstrates that high-quality reasoning (**Detailed**) improves standard tasks. However, isolated reasoning benchmarks reveal an

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

Table 4: Absolute Accuracy on downstream tasks. **Avg.** is the macro-average.

Model Variant	Bel-PL	Bel-EN	Aya-PL	Aya-EN	MoT-PL	LightR1	Avg.
PLLuM-8B-Instruct							
Original	0.609	0.656	**0.656**	0.552	0.316	**0.172**	0.494
Detailed	**0.672**	**0.742**	0.615	**0.583**	**0.374**	0.094	**0.513**
Summarized	0.623	0.673	0.572	0.486	0.352	0.070	0.463
BabyThink	0.550	0.628	0.487	0.389	0.305	0.071	0.405
Lengthy	0.512	0.557	0.364	0.279	0.309	0.070	0.349
Bielik-11B-v2.6-Instruct							
Original	0.876	0.904	0.856	**0.903**	0.624	0.521	0.781
Detailed	**0.894**	**0.937**	**0.861**	0.854	0.671	0.537	**0.792**
Summarized	0.826	0.876	0.826	0.872	0.529	0.264	0.699
BabyThink	0.810	0.871	0.757	0.867	0.497	0.266	0.678
Lengthy	0.819	0.841	0.772	0.891	**0.701**	**0.599**	0.771

expected trade-off: fine-tuning exclusively on MoT-PL boosts our target Polish reasoning but degrades English reasoning (LightR1) due to mild catastrophic forgetting of English chain-of-thought capabilities. Finally, the **Lengthy** dataset exhibits a volatile profile: on the smaller PLLuM-8B, it caused catastrophic forgetting, but in the larger Bielik-11B model, it unlocked significant reasoning capabilities, increasing performance on MoT-PL-eval by +12.3% and LightR1 by +15.0%.

Table 5: Relative Percentage Change (%) on downstream tasks between original and finetuned models.

Model	Bel-PL	Bel-EN	Aya-PL	Aya-EN	MoT-PL	LightR1	Avg.
PLLuM-8B-Instruct							
Detailed	**+10.3%**	**+13.1%**	-6.2%	**+5.6%**	**+18.4%**	-45.3%	**+3.8%**
Summarized	+2.3%	+2.6%	-12.8%	-12.0%	+11.4%	-59.3%	-6.3%
BabyThink	-9.7%	-4.3%	-25.8%	-29.5%	-3.5%	-58.7%	-18.0%
Lengthy	-15.9%	-15.1%	-44.5%	-49.5%	-2.2%	-59.3%	-29.4%
Bielik-11B-v2.6-Instruct							
Detailed	**+2.1%**	**+3.7%**	**+0.6%**	-5.4%	+7.5%	+3.1%	**+1.4%**
Summarized	-5.7%	-3.1%	-3.5%	-3.4%	-15.2%	-49.3%	-10.5%
BabyThink	-7.5%	-3.7%	-11.6%	-4.0%	-20.4%	-48.9%	-13.2%
Lengthy	-6.5%	-7.0%	-9.8%	-1.3%	**+12.3%**	**+15.0%**	-1.3%

The domain-specific breakdown in Table 6 exposes a critical dependency between model capacity and reasoning density. In **MATH**, we observe a striking inversion of preferences: the smaller **PLLuM-8B** benefits exclusively from the *Summarized* variant (+26.2%), likely succumbing to context drift in longer chains, whereas the larger **Bielik-11B** effectively utilizes the "thinking space" of *Lengthy* derivations (+12.5%) to navigate complex logic. This capacity gap is most acute in **CODE**, where verbose reasoning acts as a crucial scaffold for Bielik-11B (+131.4%) but induces catastrophic forgetting in PLLuM-8B (-73% to -96%). In contrast, in **SCIENCE**, the smaller model sees the largest relative gains (+28.6%), suggesting that reasoning traces help unlock latent knowledge, while the larger model hits a performance ceiling with only marginal improvements (+5.0%).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

Table 6: MoT-PL-eval performance by domain. Cells show **Accuracy** followed by **(Relative Gain %)** compared to the Original baseline. Bold indicates the best result per model/domain.

Model	MATH	CODE	SCIENCE
PLLuM-8B-Instruct			
Original	0.103	**0.138**	0.479
Detailed	0.100 (-2.9%)	0.020 (-85.5%)	**0.616 (+28.6%)**
Summarized	**0.130 (+26.2%)**	0.006 (-95.7%)	0.574 (+19.8%)
BabyThink	0.058 (-43.7%)	0.012 (-91.3%)	0.523 (+9.2%)
Lengthy	0.076 (-26.2%)	0.037 (-73.2%)	0.511 (+6.7%)
Bielik-11B-v2.6-Instruct			
Original	0.593	0.175	0.785
Detailed	0.577 (-2.7%)	0.346 (+97.7%)	**0.824 (+5.0%)**
Summarized	0.276 (-53.5%)	0.112 (-36.0%)	0.791 (+0.8%)
BabyThink	0.285 (-51.9%)	0.127 (-27.4%)	0.722 (-8.0%)
Lengthy	**0.667 (+12.5%)**	**0.405 (+131.4%)**	0.809 (+3.1%)

4.4 Correlation Analysis: Drivers of Reasoning Performance

To understand the mechanisms behind the observed performance changes, we analyzed the relationship between our training data metrics (defined in Sections 3.1 and 3.2) and the downstream performance. We calculated the Spearman Rank Correlation (ρ) between each metric and the relative performance gain.

Table 7 highlights a distinct divergence in how training data characteristics translate to downstream performance. For **PLLuM-8B**, performance is primarily driven by **Semantic Alignment** ($\rho_{avg} = 0.75$), **Semantic Flow** ($\rho_{avg} = 0.65$) and **Factuality** ($\rho_{avg} = 0.45$). This suggests that the smaller model relies heavily on clear, instruction-compliant data. In particular, **Utility** shows a strong negative correlation with the complex LightR1 benchmark ($\rho = -0.74$). This indicates that data optimized for high utility, typically concise summaries, deprive the model of the intermediate reasoning tokens necessary to learn complex logic steps. In contrast, **Bielik-11B** demonstrates a strong dependence on reasoning volume and correctness. Although the **Redundancy Ratio** perfectly predicts the success in LightR1 ($\rho = 1.0$), the model-based metrics clarify the nature of this redundancy. **Validity** and **Coherence** show near-perfect correlations with reasoning tasks, confirming that the model leverages redundant tokens effectively only when they form a logically valid reasoning. In contrast, **Semantic Flow** correlates negatively with hard reasoning ($\rho = -0.80$), reinforcing that narrative smoothness is less critical than a rigorous step-by-step derivation for the larger model.

Table 8 differentiates the drivers for procedural logic versus knowledge retrieval. In the Code and Math domains, **Semantic Flow** correlates negatively for Bielik-11B (reaching $\rho = -0.80$), indicating that narrative smoothness often impedes strict logical derivation. For Bielik-11B, performance in these domains depends on a combination of reasoning volume and correctness. The model shows a perfect correlation with **Redundancy Ratio** ($\rho = 1.0$) alongside strong correlations with **Validity** and **Coherence** ($\rho = 0.80$). This suggests that the benefit of verbose reasoning comes from the generation of valid and coherent intermediate

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

Table 7: Spearman's ρ between training dataset metrics and downstream performance on general benchmarks. The metrics are divided into **Analytical** (on full dataset) and **Model-based** (on a stratified subsample of 1,000 examples).

Metric	Bel-PL	Bel-EN	Aya-PL	Aya-EN	MoT-PL	LightR1	Avg.
PLLuM-8B-Instruct							
Analytical Metrics							
Redundancy Ratio	-0.40	-0.40	-0.40	-0.40	0.00	0.11	-0.25
Semantic Alignment	**0.80**	**0.80**	**0.80**	**0.80**	**1.00**	0.32	**0.75**
Semantic Flow	**0.80**	**0.80**	**0.80**	**0.80**	0.60	0.11	0.65
Symbolic Fraction	0.60	0.60	0.60	0.60	0.80	-0.21	0.50
Syntactic Depth	0.00	0.00	0.00	0.00	0.40	**-0.74**	-0.06
Model-based Metrics							
Validity (V)	-0.20	-0.20	-0.20	-0.20	0.40	-0.21	-0.10
Factuality (F)	**0.40**	**0.40**	**0.40**	**0.40**	**0.80**	0.32	**0.45**
Coherence (C)	-0.20	-0.20	-0.20	-0.20	0.40	-0.21	-0.10
Utility (U)	0.00	0.00	0.00	0.00	0.40	**-0.74**	-0.06
Bielik-11B-v2.6-Instruct							
Analytical Metrics							
Redundancy Ratio	0.00	-0.40	0.00	0.20	**0.80**	**1.00**	0.27
Semantic Alignment	**1.00**	**0.80**	**1.00**	-0.40	0.40	0.00	**0.47**
Semantic Flow	0.60	**0.80**	0.60	-0.40	-0.40	-0.80	0.07
Symbolic Fraction	0.80	0.60	0.80	0.00	0.20	-0.40	0.33
Syntactic Depth	0.40	0.00	0.40	**0.60**	0.40	-0.20	0.27
Model-based Metrics							
Validity (V)	0.40	-0.20	0.40	0.40	**1.00**	**0.80**	0.47
Factuality (F)	**0.80**	**0.40**	**0.80**	-0.20	0.80	0.60	**0.53**
Coherence (C)	0.40	-0.20	0.40	0.40	**1.00**	**0.80**	0.47
Utility (U)	0.40	0.00	0.40	**0.60**	0.40	-0.20	0.27

steps rather than redundancy alone. PLLuM-8B shows a divergent pattern in MATH, where performance correlates perfectly with **Symbolic Fraction** ($\rho = 1.0$) and strongly with **Utility** ($\rho = 0.80$), but weakly with **Validity** ($\rho = 0.20$). This implies a reliance on formal notation and concise answers rather than the verification of the logical chain. However, in CODE, PLLuM-8B aligns with the larger model, showing strong correlations with both **Redundancy** ($\rho = 1.0$) and **Validity** ($\rho = 0.80$). Finally, SCIENCE is distinct; here, Bielik-11B exhibits a perfect correlation with **Factuality** ($\rho = 1.0$), identifying factual accuracy as the sole critical driver, while PLLuM-8B relies primarily on **Semantic Flow** and **Semantic Alignment** ($\rho = 0.80$).

5 Discussion

RQ1. Is it feasible to validate the utility prior to fine-tuning? **Yes, but the predictive signal of the metrics depends the model size.** Our analysis confirms that dataset metrics are reliable performance predictors ($\rho \geq 0.75$), yet there is no universal quality profile. For example, *Redundancy Ratio* acts as a decisive positive signal for the Bielik-11B model in reasoning tasks ($\rho = 1.0$) but remains neutral or negative for the PLLuM-8B model. Similarly, while *Semantic*

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

Table 8: Spearman's ρ on **Reasoning Domains** for **MoT-PL** dataset comparison. Left side: PLLuM-8B, Right side: Bielik-11B.

	PLLuM-8B			**Bielik-11B**		
Metric	**MATH**	**CODE**	**SCIENCE**	**MATH**	**CODE**	**SCIENCE**
Analytical Metrics						
Redundancy Ratio	-0.40	**1.00**	-0.40	**1.00**	**1.00**	0.60
Semantic Alignment	0.80	0.00	**0.80**	0.00	0.00	**0.80**
Semantic Flow	0.80	-0.80	**0.80**	-0.80	-0.80	0.00
Symbolic Fraction	**1.00**	-0.40	0.60	-0.40	-0.40	0.40
Syntactic Depth	0.80	-0.20	0.00	-0.20	-0.20	0.20
Model-based Metrics						
Validity (V)	0.20	**0.80**	-0.20	**0.80**	**0.80**	0.80
Factuality (F)	0.40	0.60	**0.40**	0.60	0.60	**1.00**
Coherence (C)	0.20	**0.80**	-0.20	**0.80**	**0.80**	0.80
Utility (U)	**0.80**	-0.20	0.00	-0.20	-0.20	0.20

Alignment universally benefits general instruction following, it fails to predict success in complex reasoning for larger models. This indicates that pre-validation of training data requires a scale-based calibration; small models benefit more from semantic coherence with less redundancy, while larger models can more effectively leverage redundancy in longer reasoning trace.

RQ2. Which specific quantitative measures provide the most meaningful signal for validating training data quality? We observe a fundamental dichotomy in metric efficacy driven by the complexity threshold of the model.

For **PLLuM-8B**, performance is driven by *Semantic Alignment* ($\rho = 0.75$) and *Factuality*. However, we observe a distinct negative correlation between *Utility* and complex reasoning ($\rho = -0.74$ in LightR1). This suggests that data optimized for high human utility deprives smaller models of the intermediate tokens necessary to learn logic. Thus, for smaller models, the most critical signal is the directness and factual grounding of the data, rather than its reasoning depth.

For **Bielik-11B**, *Redundancy Ratio* is the strongest predictor of reasoning success ($\rho = 1.0$), provided that it is supported by high *Validity* ($\rho = 0.80$). Crucially, *Semantic Flow* correlates negatively with Math and Code performance ($\rho = -0.80$). This indicates that the larger model benefit from verbose, rigorous derivation steps, even if repetitive, rather than smooth narrative explanations. In knowledge-heavy domains like Science, this shifts entirely to *Factuality* ($\rho = 1.0$), rendering structural metrics less relevant.

6 Conclusions

This study establishes that effective data validation requires calibrating metrics to model capacity. By analyzing the correlation between the properties of the intrinsic data and the downstream performance, we identified distinct optimization requirements for different scales of parameters.

For a smaller model (PLLuM-8B), we observed a negative correlation between metrics favoring conciseness (Utility) and reasoning performance. These models

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

rely primarily on *Semantic Alignment* and *Factuality* to prevent hallucinations, suggesting that training data should prioritize direct instruction adherence over complex reasoning chains. In contrast, the larger model (Bielik-11B) demonstrated a strong positive correlation with *Redundancy Ratio* in formal domains. This indicates that verbose iterative derivation steps are essential for performance on this scale. Consequently, data curation must distinguish between knowledge-intensive tasks, which benefit from factual density, and reasoning tasks, which require structural redundancy.

Building on these findings, future work will focus on extending this capacity-aware validation framework across a wider spectrum of model sizes to pinpoint the exact parameter threshold for reasoning verbosity. Additionally, we aim to employ instance-level influence functions to establish a direct causal link between specific structural data patterns and inference-time logical robustness. Simultaneously, we will investigate how much the reasoning setup affects other LLM properties, such as the tendency towards hallucination [24] or in-context learning [34].

7 Limitations

Our findings suggest that 8B and 11B models use verbose reasoning data differently, but because we did not test intermediate or much larger models, we cannot tell whether this shift is gradual or appears at a specific scale. We also used disjoint train and test sets, which supports rigorous system-level correlation analysis but obscures instance-level effects, so we cannot identify how particular reasoning patterns affect individual predictions. In addition, our evaluation depends on an LLM judge, which may introduce bias despite strong agreement with human raters; reasoning variants generated with DeepSeek-V3 may contain paraphrasing artifacts that models can exploit; and testing Polish-fine-tuned models on English benchmarks introduces cross-lingual effects that make it harder to isolate reasoning ability from language processing limitations.

Acknowledgments. This work was supported by: (1) the National Science Center, Poland, grant no. 2021/41/B/ST6/04471; (2) CLARIN-PL: Common Language Resources and Technology Infrastructure (POIR.04.02.00-00C002/19, 2024/WK/01, FENG.02.04-IP.040004/24); (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) the statutory funds of the Dept.of AI, Wroclaw Tech; (5) Polish Ministry of Education and Science: "International Projects Co-Funded"; (6) the EU under the Horizon Europe, grant no. 101086321 (OMINO). The views expressed are those of the authors and do not necessarily reflect those of the EU or the European Research Executive Agency.

Disclosure of Interests. All authors have received funding from the Ministry of Digital Affairs of Poland, Polish National Science Center, and the European Union.

References

1. Bandarkar, L., et al.: The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In: ACL. pp. 749–775 (2024)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

2. Bercovich, A., et al.: Llama-nemotron: Efficient reasoning models (2025)
3. Chang, T.A., et al.: Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures. arXiv preprint arXiv:2510.24081 (2025)
4. Chen, D., et al.: Data-juicer: A one-stop data processing system for large language models. In: Proceedings of SIGMOD. pp. 120–134 (2024)
5. Chen, Y., et al.: Reasoning models don't always say what they think. arXiv preprint arXiv:2505.05410 (2025)
6. Chodak, G., et al.: Typology of image crises using large language models: A novel approach to crisis classification. J. of Contingencies and Crisis Management (2025)
7. Chua, J., Evans, O.: Are deepseek r1 and other reasoning models more faithful? In: ICLR Workshop on Foundation Models in the Wild (2025)
8. Dadas, S., et al.: PIRB: A comprehensive benchmark of Polish dense and hybrid text retrieval methods. In: Proceedings of LREC-COLING. pp. 12761–12774 (2024)
9. DeepSeek-AI: Deepseek-v3 technical report (2024)
10. Ferdinan, T., et al.: Architectural concepts for integrating fundamental drives and emotions into artificial intelligence. IEEE Intelligent Systems (2025)
11. Grattafiori, A., et al.: The llama 3 herd of models (2024)
12. Guo, D., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
13. HuggingFace: Open r1: A fully open reproduction of deepseek-r1 (2025), https://github.com/huggingface/open-r1
14. Hwang, H., et al.: Assessing LLM reasoning steps via principal knowledge grounding. In: Findings of EMNLP. pp. 19925–19948 (2025)
15. Jiang, A.Q., et al.: Mistral 7b (2023)
16. Jin, M., et al.: The impact of reasoning step length on large language models. In: Findings of ACL. pp. 1830–1842 (2024)
17. Kambhampati, S., et al.: Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! arXiv preprint arXiv:2504.09762 (2025)
18. Kocoń, J., et al.: PLLuM: A Family of Polish Large Language Models. arXiv preprint arXiv:2511.03823 (2025)
19. Langner, M., et al.: Divide, cache, conquer: Dichotomic prompting for efficient multi-label llm-based classification. In: 2025 IEEE International Conference on Data Mining Workshops (ICDMW) (2025)
20. Lawsen, A.: Comment on the illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity (2025)
21. Lee, J., Hockenmaier, J.: Evaluating step-by-step reasoning traces: A survey. In: Findings of EMNLP. pp. 1789–1814 (2025)
22. Lightman, H., et al.: Let's verify step by step. In: ICLR (2024)
23. Lozhkov, A., et al.: Openr1-math-220k. https://huggingface.co/datasets/open-r1/OpenR1-Math-220k (2025)
24. Matys, P., et al.: AggTruth: Contextual Hallucination Detection using Aggregated Attention Scores in LLMs. In: ICCS'2025. pp. 227–243. Springer (2025)
25. Ociepa, K., et al.: Bielik 11b v2 technical report (2025)
26. OpenAI: Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini (2025)
27. Penedo, G., et al.: Codeforces cots. https://huggingface.co/datasets/open-r1/codeforces-cots (2025)
28. Pęzik, P., et al.: The PLLuM Instruction Corpus. arXiv preprint arXiv:2511.17161 (2025)
29. Pihulski, D., et al.: Breaking the illusion of reasoning in Polish LLMs: Quality over quantity of thought. In: Findings of EACL. pp. 1796–1811. ACL (2026)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19

30. Qwen Team: Qwen3 technical report (2025)
31. Rae, J.W., et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 (2021)
32. Shojaee, P., et al.: The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity (2025)
33. Singh, S., et al.: Aya dataset: An open-access collection for multilingual instruction tuning. In: Proceedings of ACL (2024)
34. Szczęsny, A., et al.: Leveraging positional bias of llm in-context learning with class-few-shot and maj-min alternating ordering. In: ICCS'2025. pp. 54–62 (2025)
35. Teng, F., et al.: Atom of thoughts for markov llm test-time scaling (2025)
36. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS **35**, 24824–24837 (2022)
37. Wen, L., et al.: Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond (2025)
38. Woźniak, S., et al.: Personalized large language models. In: 2024 IEEE International Conference on Data Mining Workshops (ICDMW) (2024)
39. Wu, Y., et al.: When more is less: Understanding chain-of-thought length in llms. arXiv preprint arXiv:2502.07266 (2025)
40. Zanotto, S.E., Aroyehun, S.: Linguistic and embedding-based profiling of texts generated by humans and large language models. In: Proceedings of EMNLP (2025)

8 Appendix

Experiments were conducted on the WCSS LEM cluster[2] using nodes equipped with 4× NVIDIA H100-94GB GPUs and Intel Xeon Platinum 8462Y+ CPUs. We utilized the `trl` library with DeepSpeed ZeRO Stage-3. Table 9 details the hyperparameters for both model families. We used the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$). All model outputs were generated using fixed decoding strategies: temperature=0.6, top$-p$=0.95, top$-k$=20, min$-p$=0.1, and a repetition penalty of 1.2.

Table 9: Fine-tuning hyperparameters used in our experiments.

Model	Epochs	Batch	Seq. Len	Peak LR	Sched.	Decay
PLLuM-8B	2	128	8,192	4×10^{-5}	Cosine	0.1
Bielik-11B	3	128	8,192	7×10^{-6}	Linear	0.0

[2] https://www.wcss.pl/en/

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_19