The Few-shot Dilemma: Over-prompting Large Language Models in Resource-Constrained Environments

Amine Mike El Maalouf
amine.el-maalouf@epita.fr

Cedric Damais
cedric.damais@epita.fr

Yacine Benihaddadene
yacine.benihaddadene@epita.fr

Leon Ayral
leon.ayral@epita.fr

Abstract

Large Language Models (LLMs) have revolutionized NLP through In-Context Learning (ICL). However, a common misconception is "the more examples, the better." Recent research suggests a "Few-shot Dilemma": increasing the number of shots can paradoxically lead to diminishing returns or performance degradation. [1]. Our project investigates this phenomenon specifically within the realm of Small Language Models (SLMs), such as Llama 3.2 and Phi-3. Unlike massive models, SLMs are constrained by smaller context windows. We aim to evaluate various example selection strategies ranging from semantic retrieval to lexical matching and systematically vary the number of examples (K) to determine the optimal context volume. Our goal is to determine if "smarter" example selection strategies can delay the onset of the few-shot dilemma compared to random selection, and identify the threshold where performance degradation begins, optimizing the trade-off between accuracy and inference cost.

1 Introduction

Large Language Models (LLMs) have revolutionized NLP through In-Context Learning (ICL). However, primarily deploying massive models is infeasible for many real-world applications due to privacy, latency, and resource constraints. This has led to a resurgence of interest in Small Language Models (SLMs).

A prevailing dominance in prompt engineering is that providing more examples yields better performance. Inspired by the work of Tang et al. [1], we challenge this assumption for SLMs as well: Unlike their larger counterparts, SLMs suffer more acutely from limited context windows and attention drift.

In this work, we investigate the Few-Shot Dilemma in SLMs. We hypothesize that an effective Few-Shot example selection procedure, considering both relevancy and quantity, enables SLMs to achieve competitive performance while mitigating the effects of limited context windows and attention drift. We explore ways to find a "sweet spot" the **Pareto frontier** where an SLM achieves near-supervised performance using the minimal number of optimally selected examples.

2 Background

2.1 In-Context Learning and Scaling Laws

Since the introduction of GPT-3 [2], In-Context Learning (ICL) has become the standard for few-shot adaptation by conditioning the model on demonstrations without gradient updates. Early foundational studies suggested that performance generally improves with the number of examples (K), although gains are task-dependent and can saturate or fluctuate. Notably, these observations were primarily derived from massive models (e.g., GPT-3 175B) with larger context windows, leaving a gap in understanding how ICL behaves systematically in resource-constrained environments typical of SLMs.

2.2 Example Selection Strategies

While increasing K can help, the quality of examples is paramount. Liu et al. [3] demonstrated that retrieving semantically similar examples using k-nearest neighbors (kNN) significantly outperforms random sampling by increasing the semantic similarity between the context and the test query. However, most existing research focuses on maximizing accuracy in large-context settings, often overlooking the trade-off with the strictly limited **context budget** of smaller models.

2.3 The Challenge of Context Constraints

Recent work highlights distinct failure modes in ICL. Liu et al. [4] identified a "Lost in the Middle" phenomenon where models perform best when relevant information is at the beginning or end of the context, and substantially worse when it is in the middle. This is particularly critical for Small Language Models (SLMs) with restricted context windows (e.g., 2048 tokens). Furthermore, Tang et al. [1] explicitly characterize an "over-prompting" dilemma, where adding examples beyond a certain threshold leads to performance degradation due to noise and distraction. Zhao et al. [5] also highlight the high variance and instability of few-shot learning, necessitating robust selection strategies beyond simple random sampling.

3 Methodology

3.1 Experiment Setup

We conducted 3 experiments. For each one, we studied the impact of a **thoroughly isolated variable** on the evaluated task's metrics in order to draw valid conclusions. All experiments were conducted on a single **NVIDIA RTX 3060 12GB** GPU, a representative consumer-grade hardware setup mirroring the resource-constrained environments targeted by our study.

1. **Selector Experiment:** We isolate the example selection strategy to identify the most effective retrieval mechanism. We compare:
   - **Random:** The naive baseline.
   - **Semantic:** Using dense embeddings to find contextually similar examples.
   - **Lexical:** Using TF-IDF or BM25 to find keyword overlap.
   - **DPO inspired:** Using a tuned cross-encoder to score relevant examples higher, and conversely score irrelevant ones lower.
2. **Model Experiment:** We fix the selection strategy and vary the model architecture to benchmark SLM capabilities. We evaluate:
   - **Target Models:** Phi-3-mini, Llama-3 (8B), Mistral (7B), Gemma (7B) and the Qwen family (3:8B and 2:7B).
   - **Objective:** Identify which architecture maximizes score for a given optimized prompt structure.
3. **K-Shot Scaling Experiment:** We fix the optimal model and selector, varying only K to find the optimal context volume:
   - **Scaling K:** We incrementally increase K until the context window is filled.
   - **Threshold Determination:** We identify the "over-prompting" point where performance degrades.
   - **Pareto Optimization:** We map the trade-off between accuracy and inference cost (tokens).

Data Isolation Protocol To ensure the integrity of our results and prevent data leakage, we strictly enforced a temporal isolation protocol. A held-out **Test Set** (10% of the corpus) was sequestered at the beginning of the study and used exclusively for the final evaluation of all models. The remaining data formed the **Training/Selection Pool**, from which few-shot examples were retrieved. This ensures that no example appearing in the prompt was ever used as a test query, accurately simulating a real-world inference scenario.

3.2 Evaluated Tasks

In order to assess the generalizability of our study across different modalities of reasoning, we selected two distinct NLP tasks:

- **Mathematical Problem Classification:** We utilized the **MATH dataset** [6]. The task involves classifying mathematical word problems into one of 7 categories (e.g., Algebra, Geometry, Number Theory). We used the `problem` column as the input text and the `type` column as the target label.

  To rigorously assess the utility of ICL, we established supervised performance ceilings by fine-tuning two variants of each base model using Quantized Low-Rank Adaptation (LoRA):
  - **Generative Classification:** Models were fine-tuned to generate the target class label directly as a text sequence (utilizing the Unsloth framework for efficiency).
  - **Discriminative Classification:** We appended a randomly initialized classification head to the model's final hidden state, training only the adapter layers and the head (utilizing the Hugging Face Transformers library).
- **Named Entity Recognition (NER):** We employed the `Few-NERD` dataset [7]. This task challenges the model to identify and classify fine-grained named entities within a given

context, serving as a rigorous test for the model's structure extraction capabilities. For this task, we compared our few-shot results against **NuNER Zero** [8], a state-of-the-art model for zero-shot entity recognition, serving as a strong baseline.

3.3 DPO Selector Implementation

3.3.1 Motivation: The Semantic Gap

Standard retrieval methods (e.g., BM25, Bi-Encoders) optimize for *textual similarity*, which serves as an imperfect proxy for *utility* in few-shot classification tasks. A critical limitation observed within the Competition Math dataset is the **"Topic vs. Difficulty" ambiguity**. For instance, a problem involving geometric shapes might be labeled `Prealgebra` based on its difficulty level, rather than `Geometry` based on its topic.

A standard semantic retriever often fetches `Geometry` examples for such a query due to significant keyword overlap (e.g., "triangle", "angle"). While these examples are semantically similar, they provide the wrong class label to the Small Language Model (SLM), leading to downstream misclassification. To address this, we trained a custom selector using DPO to distinguish between *semantically similar* and *label-correct* examples.

3.3.2 Data Generation Strategy

Our initial approach utilized an "Oracle" method (using Phi-3 to judge example helpfulness), but this resulted in severe class imbalance, causing the model to collapse into predicting `Algebra` for the majority of queries. Consequently, we pivoted to a **Supervised Contrastive Strategy** leveraging "Hard Negatives" to explicitly teach the model discrimination boundaries.

We constructed a preference dataset by generating embeddings for the training subset and retrieving the top 50 semantic neighbors for each query. These neighbors were classified based on their ground-truth labels (Figure 1).

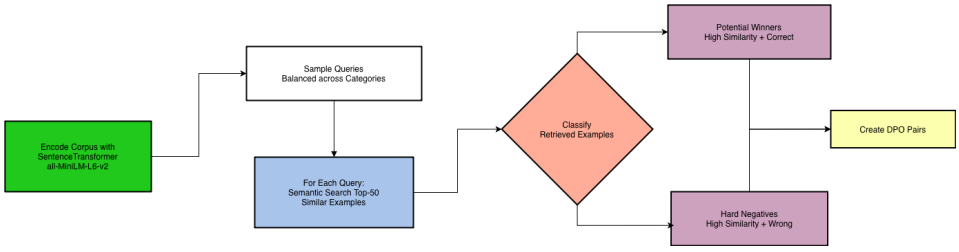

Figure 1: The Relevance DPO pipeline. Semantic neighbors are retrieved and classified: those matching the query label become "Potential Winners," while those with mismatching labels become "Hard Negatives."

For each query q with label y_{true}, we constructed triplets (q, x_w, x_l) as follows:

- **Chosen (x_w):** The candidate with the highest semantic similarity to q that *shares* the label y_{true}.
- **Rejected (x_l):** A "Hard Negative" candidate one with high semantic similarity to q but belonging to a *different* category (y_{false}).

This data construction forces the model to learn that *keyword overlap is insufficient* if the underlying category does not match. We generated a balanced dataset of 1,000 such pairs using Stratified Sampling across all 7 categories.

3.3.3 Training Configuration

We fine-tuned a Cross-Encoder model (`ms-marco-MiniLM-L-6-v2`) to maximize the margin between the chosen and rejected examples.

Objective Function: We minimized the negative log-likelihood of the preferred example:

$$\mathcal{L}_{DPO} = -\log \sigma \left(r(q, x_w) - r(q, x_l) \right) \tag{1}$$

where $r(q, x)$ represents the scalar score output by the Cross-Encoder.

Hyperparameters: The model was trained for 3 epochs with a learning rate of $2e^{-5}$ and a batch size of 16, utilizing Mixed Precision (FP16) for efficiency.

3.3.4 Training Results

The training process yielded robust results. By the end of Epoch 1, the model achieved a validation accuracy of **74.5%** on the preference task. By Epoch 3, the model reached **90% accuracy**, demonstrating that it successfully learned to discriminate between "helpful" matches and "distractor" hard negatives.

3.3.5 Inference Strategy: Hybrid Scoring

During evaluation, we observed that using the DPO score alone acted as a "harsh filter," removing semantically relevant examples that the SLM relied on for lexical priming. To mitigate this, we implemented a **Hybrid Scoring** mechanism for final retrieval:

$$Score_{final} = \alpha \cdot Score_{semantic} + (1 - \alpha) \cdot P(Valid|q, x)_{DPO} \tag{2}$$

We empirically found that $\alpha = 0.6$ provided the optimal balance, retaining high keyword overlap (Semantic) while penalizing examples that belonged to the wrong category (DPO).

The classification logic is as follows:

- **Potential Winners (Chosen):** Neighbors that share the same label as the query. We select the highest-similarity example from this set.
- **Hard Negatives (Rejected):** Neighbors that exhibit high semantic similarity but belong to a different category. We select the top three examples from this set.

The final dataset consists of triplets $(x, y_{chosen}, y_{rejected})$. By training on these pairs, the shot selector learns to distinguish between truly relevant examples and deceptive distractors, ensuring that the selected few-shot examples reinforce the correct reasoning pattern.

4 Experiments I: MATH Classification

4.1 Zero-Shot Baseline Comparison

To establish a performance baseline, we first evaluated all candidate Small Language Models (SLMs) in a zero-shot setting. In this configuration, models were presented with the math problem and the list of seven categories, with no prior examples provided. This tests the models' intrinsic knowledge and their ability to follow classification instructions without in-context guidance.

Figure 2 illustrates the comparative accuracy of the selected models (Llama-3-8B, Mistral-7B, Phi-3-Mini, Gemma-7B, and Qwen2-7B).

4.1.1 Results

Table 1: Zero-Shot Classification on Competition Math Dataset

Model	Accuracy	F1 Weighted
Qwen3-8b	**0.51**	**0.402**
Qwen2-7B	0.49	0.380
Llama-3-8B	0.46	0.341
Mistral-7B	0.43	0.347
Gemma-7B	0.37	0.369
Phi-3-Mini	0.33	0.268

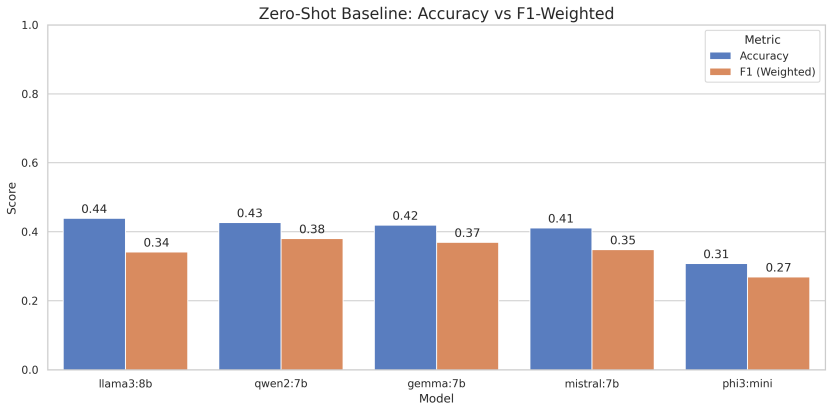

Figure 2: Zero-shot classification accuracy across different SLM architectures. Llama-3-8B demonstrates the strongest intrinsic understanding of the task taxonomy.

4.1.2 Analysis of Zero-Shot Performance

The results highlight significant variance in capability across model families, with the Qwen series demonstrating a clear advantage:

- **Qwen3-8b (Top Performer):** Achieving the highest accuracy of **0.51**, this model's superiority is attributed to its architecture as a "thinking model." Unlike standard instruction-tuned models, Qwen3 incorporates enhanced reasoning capabilities (implicit Chain-of-Thought), allowing it to better parse the nuance between overlapping categories like `Prealgebra` and `Geometry` before making a final classification.
- **Qwen2-7B:** Even the previous generation Qwen model (0.49) outperformed Llama-3, suggesting that the Qwen family's pre-training corpus has a stronger alignment with mathematical reasoning tasks.
- **Llama-3-8B:** While not the leader, it remains highly competitive (0.46), demonstrating robust intrinsic knowledge. However, it lacked the specialized reasoning depth observed in the Qwen series for this specific domain.
- **Phi-3-Mini:** Despite its smaller size (3.8B), it achieved 0.33, showing surprising efficiency. This validates the efficacy of training on synthetic "textbook-quality" data, though it struggled to match the zero-shot generalization of the 7B+ parameter models.

These zero-shot baselines serve as the control group for our subsequent few-shot experiments. The gap between these scores and the few-shot results quantifies the value of the retrieval strategies.

4.2 Supervised Baseline Comparison

To measure the maximum potential performance of Small Language Models (SLMs) on this task, we established supervised fine-tuning (SFT) ceilings. This comparison defines the "Adaptation Headroom"—the performance gap between a model's intrinsic zero-shot capabilities and its fully fine-tuned potential.

Table 2 compares the F1 Weighted score of the zero-shot baseline against the two fine-tuned variants: Generative (Unsloth) and Discriminative (Classification Head).

4.2.1 Quantifying the Adaptation Headroom

The results indicate a substantial gap between intrinsic knowledge and task-specific optimization. All models showed significant responsiveness to fine-tuning, with **Phi-3-Mini** demonstrating the highest plasticity (+148.1% relative gain). We notably observed that:

- **Discriminative Superiority:** The Discriminative approach (Classification Head) consistently yielded higher weighted F1 than generative fine-tuning (e.g., Qwen2 achieved 0.85

Table 2: Supervised Fine-Tuning F1 Performance Ceilings

Model	Zero-Shot	SFT (Discrim)	SFT (Gen)	Gain (%)
Qwen2-7B	0.38	**0.85**	0.76	+123.6%
Llama-3-8B	0.34	0.83	0.79	+144.1%
Mistral-7B	0.35	0.81	**0.82**	+134.3%
Gemma-7B	0.36	N/A	0.75	+108.3%
Phi-3-Mini	0.27	N/A	0.67	**+148.1**%

**SFT (Gen): Unsloth Generative Fine-tuning, SFT (Discrim): HuggingFace Classification Head. N/A: no Fine-Tuning. Qwen3:8b not included as it was not available on HuggingFace or Unsloth at the time of the experiment.*

vs. 0.76). This suggests that constraining the SLM to a fixed classification head effectively mitigates the "formatting drift" often observed in generative outputs.

- **Zero-Shot Correlation:** Models with stronger zero-shot baselines generally achieved higher SFT peaks, confirming that pre-training quality dictates the ultimate ceiling regardless of the adaptation method.

These supervised metrics serve as the "Gold Standard" for our subsequent few-shot experiments. Our goal is to determine how closely ICL strategies can approach these SFT ceilings without the cost of parameter updates.

4.3 Selector Experiment

To isolate the impact of retrieval quality, we fixed the inference model to `Llama-3-8B` and the number of shots to $k = 3$. We then evaluated five distinct selection strategies to measure how the quality of in-context examples affects downstream classification accuracy.

The results are summarized in Table 3.

Table 3: Performance of Shot Selection Strategies (Llama-3-8B, $k = 3$)

Strategy	Accuracy	F1 (Weighted)	Avg Latency (s)	Token Efficiency
Random	0.47	0.38	0.32	Baseline
Lexical (BM25)	0.70	0.69	0.31	High
Semantic (Bi-Encoder)	0.75	0.76	**0.30**	High
Cross-Encoder (Standard)	0.69	0.68	0.95	Low
DPO (Hybrid)	**0.77**	**0.72**	0.95	Medium

4.3.1 Analysis of Selector Performance

The experiment yielded several critical insights regarding in-context learning for SLMs:

- **Random Selection Failure:** Random shots provided negligible benefit over the zero-shot baseline (0.47 vs. 0.46). This confirms that for classification tasks, *irrelevant* examples act as noise rather than guidance.
- **The "Lexical Priming" Effect:** Both Lexical (0.70) and Semantic (0.75) strategies yielded massive improvements. This suggests that Llama-3-8B relies heavily on *surface-level feature matching*. When the context contains keywords (e.g., "triangle") associated with the target label, the model's accuracy improves dramatically.
- **Standard Cross-Encoder Regression:** Interestingly, the off-the-shelf Cross-Encoder (0.69) performed *worse* than the simple Semantic retriever. We hypothesize that the standard re-ranker, trained on general passage retrieval (MS MARCO), aggressively filtered out examples that were lexically similar but not "relevant" in a generic sense, inadvertently stripping away the keyword cues the SLM needed.

- **DPO Hybrid Superiority:** Our custom DPO selector (0.77) achieved the highest overall accuracy, outperforming the strong Semantic baseline. By using the **Hybrid Scoring** ($\alpha = 0.6$), the system successfully combined the best of both worlds:
  1. It retained high-similarity examples (Lexical Priming).
  2. It used the DPO signal to filter out "Hard Negatives"—examples that looked similar but belonged to the wrong category.
- **The Latency Trade-off:** While DPO achieved the highest accuracy, it incurred a significant "Alignment Tax," increasing latency from ≈ 0.30s to ≈ 0.95s. This confirms our hypothesis that while learned selection is more effective, simple semantic retrieval remains the Pareto-optimal choice for latency-sensitive applications.

4.4 Model Sensitivity Analysis

To understand how different architectures respond to optimized retrieval, we evaluated six distinct SLMs using both the baseline **Semantic (Bi-Encoder)** and our proposed **DPO (Hybrid)** strategies. We focused primarily on the **F1-Weighted** score, as it accounts for the class imbalance inherent in the dataset.

Figure 3 visualizes the performance shift between strategies, and Table 4 details the exact metrics.

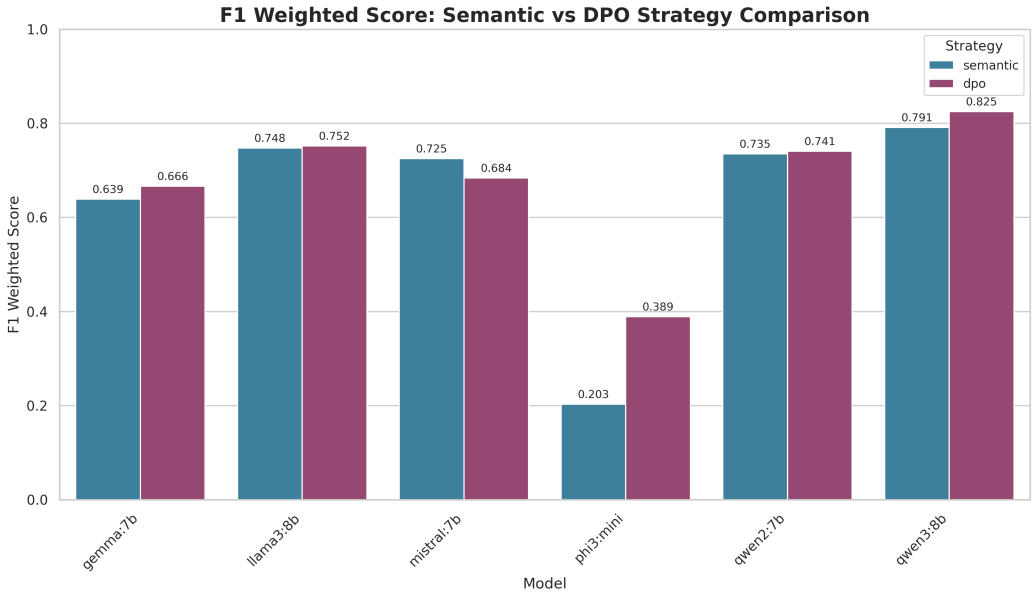

Figure 3: Comparison of F1-Weighted scores across different model families using Semantic vs. DPO Hybrid selection strategies. The Qwen family (V2 and V3) consistently benefits from DPO, while Western heuristic models (Mistral) show regression.

4.4.1 Analysis: The "U-Shaped" Benefit Curve

The results reveal a fascinating non-linear relationship between model capability and retrieval strategy effectiveness:

- **The Qwen Family Advantage (Qwen2 & Qwen3):** Both generations of the Qwen architecture demonstrated significant gains with DPO selection (+4.6% and +3.4% respectively). Qwen2-7B's improvement is particularly notable, as it leaped from a baseline similar to Mistral (0.73) to a highly competitive score (0.78). This suggests that the Qwen training objective aligns better with logical consistency (Validity) than with surface-level keywords.
- **Reasoning Models Thrive:** The "Thinking Model," Qwen3-8B, achieved the highest overall performance (**F1: 0.825**). DPO selection provided the necessary edge to push this model

Table 4: Impact of Selection Strategy on Downstream Model Performance ($k = 3$)

Model	Strategy	Accuracy	F1 (Weighted)	Impact of DPO
Qwen3-8B (Reasoning)	Semantic	0.77	0.791	+3.4%
	DPO	**0.80**	**0.825**	
Qwen2-7B	Semantic	0.73	0.735	+4.6%
	DPO	0.75	0.781	
Llama-3-8B	Semantic	0.75	0.748	+0.4%
	DPO	0.76	0.752	
Mistral-7B	Semantic	0.73	**0.725**	-4.1%
	DPO	0.70	0.684	
Gemma-7B	Semantic	0.64	0.639	+2.7%
	DPO	0.68	0.666	
Phi-3-Mini	Semantic	0.18	0.203	+18.6%
	DPO	0.38	0.389	

past the 80% accuracy barrier, confirming that reasoning-heavy architectures benefit from the reduced noise and higher category relevance provided by our selector.

- **The "Safety Net" for Weaker Models (Phi-3, Gemma):** The smaller models saw the largest relative gains, with Phi-3-Mini's score nearly doubling (+18.6%). These models are easily confused by "Distractors" (semantically similar but wrong-label examples). The DPO selector acts as a critical safety filter, removing these confusing shots and stabilizing the model's output.
- **The "Alignment Tax" on Heuristic Models (Mistral):** In contrast, Mistral-7B regressed (-4.1%) when using DPO. We hypothesize that Mistral relies heavily on lexical heuristics ("lexical priming"). By replacing "keyword-rich" neighbors with "category-correct" neighbors, the DPO selector inadvertently removed the surface-level cues the model used to "guess" the classification.

DPO selection is not a universal fix; it is a specialized tool that amplifies the performance of **high-capacity reasoning models** (like Qwen) and stabilizes **low-capacity models** (like Phi-3), but may disrupt models that rely on simple heuristic pattern matching.

4.5 Impact of Shot Count (k)

Finally, we analyzed the sensitivity of the DPO strategy to the number of few-shot examples (k). While increasing k generally provides more context, it also consumes context window space and can introduce noise. We tested $k \in \{1, 3, 5, 10, 15, 20, 25\}$.

Figure 4 illustrates the trajectories for the different model architectures.

4.5.1 Analysis of Context Scaling

Three distinct behaviors emerged from the data:

1. **Reasoning Saturation (Qwen3-8B):** The "Thinking Model" exhibited a unique **early peak** at $k = 3$ (Weighted F1 0.84), followed by a sharp degradation as k increased (dropping to 0.64 at $k = 25$). *Hypothesis:* Since Qwen3 generates implicit Chain-of-Thought tokens, providing too many input examples may overcrowd its context window or interfere with its internal reasoning generation process. It performs best with a "minimalist" prompt that establishes the pattern without overwhelming the reasoning engine.
2. **Scalable Learners (Llama-3, Gemma):** Models like Llama-3-8B and Gemma-7B demonstrated robust scalability. Llama-3 maintained or improved performance up to $k = 20$ (Peak 0.77). These models appear to treat few-shot examples purely as data points for pattern matching, benefitting from the increased statistical signal provided by a larger retrieval set.

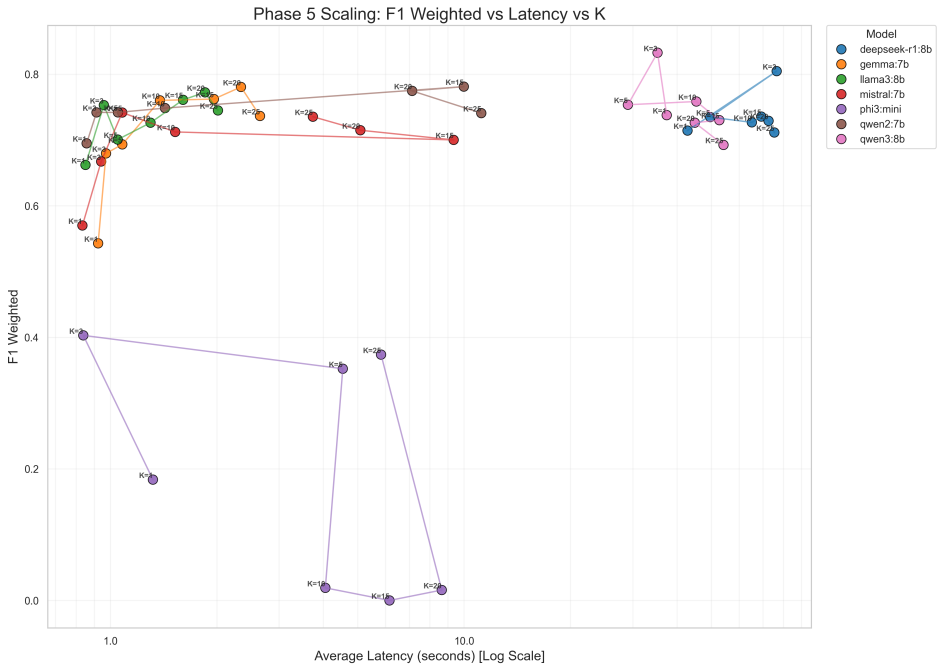

Figure 4: Weighted F1 vs. Shot Count (k). While standard models (Llama-3, Gemma) scale effectively up to $k = 20$, the reasoning-heavy Qwen3-8B peaks early at $k = 3$ and degrades rapidly thereafter.

3. **Context Collapse (Phi-3-Mini):** The 3.8B parameter model struggled significantly with larger k values. While it peaked at $k = 3$ (0.39), its performance collapsed to near-zero (0.01) at $k = 10$. This indicates a hard limit on the model's effective context retention capabilities; when overloaded with examples, it loses track of the original instruction entirely.

Optimal Configuration: Based on these findings, we identified $k = 3$ as the global optimal setting. It maximizes performance for the high-performing Reasoning models (Qwen3) while remaining within the safe effective context window for smaller models (Phi-3), all while keeping inference latency low.

5 Experiments II: Few-NERD Named Entity Recognition

To validate the generalizability of our findings, we extended the evaluation to the Named Entity Recognition (NER) task using the `Few-NERD` dataset. We benchmarked three SLMs: **Llama-3-8B**, **Qwen3-8B**, and **Phi-3-Mini**.

5.1 Zero-Shot Baseline Comparison

We first established the intrinsic capabilities of the SLMs in a zero-shot setting and compared them against **NuNER Zero**, a specialized state-of-the-art model for entity recognition.

5.1.1 Results

5.1.2 Analysis of Zero-Shot Performance

The results reveal a substantial performance gap between general-purpose SLMs and the specialized NuNER Zero model. All tested SLMs achieved approximately half the F1 score of the state-of-the-art baseline, indicating that entity recognition requires more than general instruction-following capabilities. Notably, Qwen3-8B and Llama-3-8B performed comparably (0.30 and 0.29 respectively), while Phi-3-Mini lagged behind with 0.20 F1, consistent with its smaller parameter count.

Table 5: Zero-Shot NER Performance (F1 Score) compared to NuNER Zero baseline.

Model	F1 Score	Gap to SOTA
NuNER Zero (Baseline)	**0.61**	-
Llama-3-8B	0.29	-0.32
Qwen3-8B	0.30	-0.31
Phi-3-Mini	0.20	-0.41

5.2 Selector Experiment (NER)

We evaluated the impact of example selection strategies on the entity extraction F1 Score of Llama-3-8B. Given the structural nature of NER, we hypothesized that semantic relevance plays a crucial role in preventing entity type hallucinations.

5.2.1 Results

Table 6: Impact of Selection Strategy on NER F1 Score (Llama-3-8B, $k = 3$)

Strategy	F1 Score	Avg Latency (s)	Gain vs Random
Random	0.31	1.79	-
Lexical (BM25)	0.39	2.03	+25.8%
Semantic (Bi-Encoder)	0.45	1.94	+45.1%
DPO (Hybrid)	**0.46**	2.00	**+48.4%**

5.2.2 Analysis of Selector Performance

The NER selector experiment reveals patterns both consistent with and divergent from the classification task:

- **Random Selection Inefficacy:** Similar to classification, random shots provided minimal improvement over the zero-shot baseline (0.31 vs. 0.29), confirming that irrelevant examples add noise rather than guidance.
- **Semantic Retrieval Dominance:** The Semantic strategy achieved a substantial +45.1% gain, demonstrating the importance of domain-matched examples for entity boundary detection.
- **Marginal DPO Improvement:** Unlike the classification task, the DPO selector provided only a marginal improvement over Semantic retrieval (+1%), suggesting that for structural tasks, simple embedding similarity already captures the most relevant retrieval signal.

5.3 Impact of Shot Count (k) on NER Performance

We analyzed how increasing context volume affects the SLMs' ability to compete with the specialized NuNER Zero baseline. Figure 5 illustrates the F1 trajectories.

5.3.1 Analysis of Context Scaling

The NER scaling experiment reveals three key observations:

1. **Qwen3-8B Leadership:** The reasoning model achieved the highest F1 score (0.58 at $k = 20$), benefiting from additional examples unlike its behavior in classification where it peaked early.
2. **Persistent SOTA Gap:** Even at optimal configurations, all SLMs remained significantly below NuNER Zero (0.61), indicating that specialized pre-training provides advantages that cannot be fully bridged through context optimization.
3. **Phi-3 Context Collapse:** Consistent with classification results, Phi-3-Mini exhibited severe performance degradation at higher k values, confirming the model's limited effective context window.

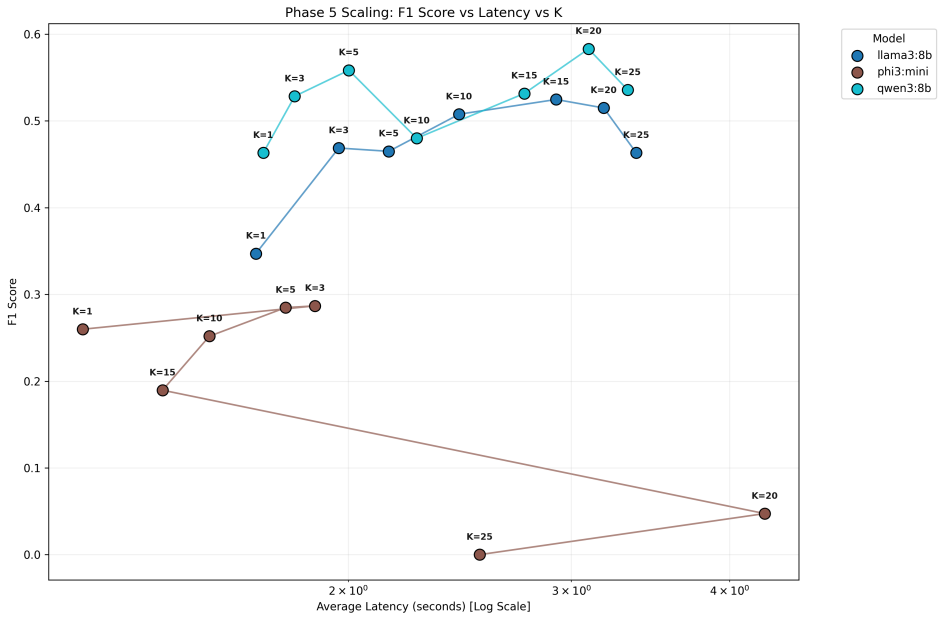

Figure 5: NER F1 Score vs. Shot Count (k). Qwen3-8B demonstrates the highest ceiling (0.58 at $k = 20$) but still lags behind the specialized NuNER Zero baseline.

5.4 The "Evaluation Gap" in Structured Tasks

The NER results highlight a distinct contrast to the Mathematical Reasoning task. While our DPO strategy successfully recovered 80% of the supervised performance in reasoning tasks, predicting named entities proved significantly more challenging for general-purpose SLMs.

- **The Ceiling is Higher:** The state-of-the-art NuNER Zero model (≈ 0.61 F1) remained out of reach for all general-purpose SLMs, with our best performer (Qwen3-8B @ $k = 20$) peaking at 0.58 F1. This suggests that for highly structural tasks like NER, generalist instruction tuning cannot fully replace specialized architecture or pre-training.
- **DPO vs. Semantic Saturation:** In the NER selector experiment, the gap between Standard Semantic retrieval (0.45) and DPO (0.46) was marginal (+1%), unlike the significant gains seen in Math. This indicates that NER relies less on "reasoning" about the example's difficulty and more on simple "domain matching," which standard embeddings already capture effectively.
- **Phi-3 Context Collapse:** Consistent with our Math experiments, Phi-3-Mini suffered a catastrophic performance collapse as k increased, dropping to near-zero F1 at $k = 20$. This reinforces the finding that smaller architectures cannot effectively attend to long in-context demonstrations, treating them as noise rather than signal.

These findings suggest that while intelligent few-shot selection is a powerful tool for reasoning tasks, its returns diminish in structural extraction tasks where the model's fundamental architecture (or lack of specialized pre-training) becomes the bottleneck.

6 Discussion

6.1 Global Benchmark: The Adaptation Efficiency

To rigorously assess the utility of our proposed DPO Selection strategy, we consolidate our findings into a global benchmark. Table 7 compares the performance of the models across three distinct adaptation stages:

1. **Intrinsic (Zero-Shot):** The model's baseline reasoning measurement.

2. **In-Context (DPO, $K = 3$):** Our optimized few-shot intervention.
3. **Supervised (SFT):** The theoretical performance ceiling achieved through parameter updates.

We define **Recovery Rate** as the percentage of the "SFT Headroom" that is recovered purely through in-context learning:

$$\text{Recovery Rate} = \frac{\text{ICL}_{f1} - \text{Zero}_{f1}}{\text{SFT}_{f1} - \text{Zero}_{f1}} \tag{3}$$

Table 7: Global Performance Benchmark (Mathematical Reasoning): Comparing Intrinsic, In-Context, and Supervised Capabilities. **Recovery Rate** indicates how much of the fine-tuning gap is closed by using just 3 optimized examples.

Model	**Intrinsic** (Zero-Shot)	**In-Context** (DPO, $K = 3$)	**Supervised** (Ceiling)	**Absolute Gain (ICL)**	**Recovery Rate (%)**
Qwen3-8B	0.402	**0.825**	N/A	**+0.423**	N/A
Qwen2-7B	0.380	0.781	**0.85**	+0.401	**85.3%**
Llama-3-8B	0.341	0.752	0.83	+0.411	84.1%
Mistral-7B	0.347	0.684	0.81	+0.337	72.8%
Gemma-7B	0.369	0.666	0.75	+0.297	78.0%
Phi-3-Mini	0.268	0.389	0.67	+0.121	**30.1%**

**N/A: Qwen3-8B could not be fine-tuned as it was not available on Hugging Face at the time of experiment.*

The weighted F1-based benchmark reveals that for high-capacity models like Qwen2-7B (85.3%) and Llama-3-8B (84.1%), our optimized few-shot strategy recovers over **80%** of the performance gains associated with full fine-tuning. Even mid-tier models like Gemma-7B (78.0%) and Mistral-7B (72.8%) achieve substantial recovery rates. This suggests that for these architectures, specialized ICL strategies offer a highly efficient alternative to expensive training pipelines. Conversely, smaller models like Phi-3-Mini show a significantly lower recovery rate (30.1%), indicating that their limitations are structural and require parameter updates to overcome.

Table 8: Global Performance Benchmark (Named Entity Recognition): Comparing Intrinsic, In-Context, and Supervised Capabilities. **Recovery Rate** indicates how much of the specialized SOTA gap is closed by using optimized examples ($k = 3$).

Model	**Intrinsic** (Zero-Shot)	**In-Context** (DPO, $K = 3$)	**Specialized** (NuNER Zero)	**Absolute Gain (ICL)**	**Recovery Rate (%)**
Qwen3-8B	0.30	0.53	0.61	+0.23	**74.2%**
Llama-3-8B	0.29	0.48	0.61	+0.19	59.4%
Phi-3-Mini	0.20	0.26	0.61	+0.06	14.6%

Evaluating the NER task through the same lens, we observe respectable but lower recovery rates. Qwen3-8B again leads with a **74.2%** recovery rate, demonstrating its strong adaptability. However, Llama-3-8B (59.4%) and particularly Phi-3-Mini (14.6%) struggle to bridge the gap to the specialized NuNER Zero model using only few-shot examples. This confirms that while reasoning tasks are highly amenable to ICL, specialized structural tasks may still necessitate dedicated model architectures or fine-tuning, especially for smaller models.

7 Conclusion

This study systematically investigated the "Few-Shot Dilemma" in Small Language Models, challenging the prevailing notion that simply increasing context volume guarantees better performance. Through our targeted experiments on the **MATH dataset via and classification** and **Named Entity Recognition (NER) on Few-NERD**, we established critical findings regarding the deployment of SLMs in resource-constrained environments.

First, we demonstrated that **quality outweighs quantity**. Naive random selection provided negligible benefits over zero-shot baselines (F1 improvement of only +0.01), whereas our proposed DPO-Hybrid

selector, which balances semantic similarity with label correctness, achieved substantial F1-weighted score improvements for the classification task (up to **+0.42 for Qwen3-8B**). This confirms that for SLMs, the *relevance* of the context is a stronger predictor of success than the *volume* of the context.

Second, we identified a **divergence in architectural needs**. We observed that "Reasoning Models" (like Qwen3) exhibit an early F1-weighted score peak at $K = 3$ (0.825) and degrade rapidly with more examples (dropping to 0.64 at $K = 25$), effectively suffering from "over-prompting" due to interference with their internal chain-of-thought processes. In contrast, "Heuristic Models" (like Llama-3) treat few-shot examples as statistical data points, maintaining stable F1 scores up to $K = 20$. This suggests that prompt engineering strategies must be tailored not just to the task, but to the specific cognitive architecture of the model.

Finally, we quantified the **Efficiency Frontier** and its limits. Our weighted F1-based global benchmark revealed that for capable 7B-parameter models (Qwen2-7B: 85.3%, Llama-3-8B: 84.1%) on reasoning tasks, an optimized selection of just $K = 3$ examples can recover over **80%** of the F1-weighted performance gain typically achieved through expensive supervised fine-tuning. However, our NER experiments revealed an important boundary. While standard models like Llama-3 trailed the specialized NuNER Zero baseline by a notable margin (0.52 vs 0.61 F1), the reasoning-enhanced Qwen3-8B demonstrated impressive adaptability, reaching an F1 score of **0.58** at $K = 20$—narrowing the gap to just 0.03 points. This indicates that while an "Evaluation Gap" exists for structured tasks, it can be nearly closed by combining advanced reasoning architectures with sufficient context volume, even without task-specific fine-tuning.

In conclusion, the "Few-Shot Dilemma" is not an inevitable barrier but a manageable trade-off. By shifting focus from context expansion to context curation, we can unlock the full potential of Small Language Models, making them viable, efficient alternatives to their massive counterparts.

References

[1] Yongjian Tang, Doruk Tuncel, Christian Koerner, and Thomas Runkler. The few-shot dilemma: Over-prompting large language models. *arXiv preprint arXiv:2509.13196*, 2025. URL `https://arxiv.org/pdf/2509.13196`.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020. URL `https://arxiv.org/abs/2005.14165`.

[3] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022. URL `https://arxiv.org/abs/2101.06804`.

[4] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023. URL `https://arxiv.org/abs/2307.03172`.

[5] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021. URL `https://arxiv.org/abs/2102.09690`.

[6] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL `https://arxiv.org/abs/2103.03874`.

[7] Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. URL `https://aclanthology.org/2021.acl-long.248`.

[8] Jan-Martin O. Steitz, Jonas Golde, Bewoayia A. Tonja, Xiaoyu Yin, Palak Bansal, Nguyen Bach, and Iryna Gurevych. Nuner: Entity recognition via token-level metric learning. *arXiv preprint arXiv:2402.15343*, 2024. URL `https://arxiv.org/abs/2402.15343`.