Mapping Researchers to Strategic Themes with Language Models via Reasoning- and Reward-Guided Fine-Tuning*

Augusto C. D. Miranda[1], Marcos P. S. Gôlo[2], Marcela A. A. dos Santos[2], Ricardo M. Marcacini[2], Marcelo A. S. Turine[1], Mariana C. de Souza[1]

[1]Faculty of Computing, Federal University of Mato Grosso do Sul

[2]Institute of Mathematics and Computer Science, University of São Paulo

{augusto.dalal,marcelo.turine,mariana.caravanti}@ufms.br

marcelaaniceto@gmail.com, {ricardo.marcacini,marcosgolo}@usp.br

Abstract. *Mapping researcher profiles to strategic topics is a key challenge in the Brazilian National Graduate System (SNPG), as it supports public policy planning and research funding allocation. To address this problem, this paper proposes TORNADO, a* ***T****w****O****-stage* ***R****easo****N****ing-rew****A****r****D*** ***O****ptimization method that combines reasoning supervision and reward optimization. Its main contribution is a distance-aware reward that models the ordinal structure of affinity labels by penalizing severe errors more than adjacent-class mistakes. Experiments on the LeanDL-HPC dataset show that TORNADO outperforms zero-shot, few-shot, and conventional supervised fine-tuning baselines, providing an efficient and scalable alternative for strategic decision-making within the SNPG.*

1. Introduction

The Brazilian National Graduate System (SNPG), coordinated primarily by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), has played a central role in structuring, evaluating, and fostering graduate education in Brazil since the 1970s [CAPES 2023]. One of the current challenges faced by the SNPG is the identification and mapping of researcher profiles to priority strategic themes defined across the 27 Federative Units of the country [Turine et al. 2025]. These themes[1] encompass areas such as sustainable development, public policy, entrepreneurship, and artificial intelligence, and are closely aligned with global initiatives, including the United Nations Sustainable Development Goals (SDGs) [United Nations 2015].

Mapping researchers to these strategic themes can support evidence-based policy-making, improve researcher allocation, and guide funding decisions [Turine et al. 2025]. To address this challenge, the LeanDL-HPC Workshop introduced a shared task focused on the accurate and explainable classification of researcher profiles and strategic themes, releasing a dataset of approximately 42,000 instances composed of theses, dissertations, and academic profile information. Recent studies have shown that proprietary Large Language Models (LLMs) can achieve strong performance in this task while also providing interpretable justifications for their predictions [Matsubura et al. 2025]. Nevertheless, such approaches remain computationally expensive and memory-intensive. Furthermore, existing solutions rely predominantly on conventional supervised fine-tuning

*Supported by FAPEC [23104.003594/2025-12] and CNPq [317002/2026-0, 408228/2025-3].

[1]https://agendanacional.capes.gov.br

[Sakiyama et al. 2025], with limited investigation of alternative paradigms, such as reasoning supervision and reward-based optimization techniques.

In this context, this paper proposes TORNADO, a **T**w**O**-stage **R**easo**N**ingrew**A**r**D** **O**ptimization method based on the fine-tuning of LLMs. TORNADO is a distance-aware reward formulation that explicitly models the ordinal structure of affinity labels by penalizing severe errors between distant classes more strongly than mistakes between adjacent classes. The framework combines supervised reasoning, learning, and reward optimization to incorporate this task-specific reward into LLM training. Our hypothesis is that explicitly encoding ordinal information in the reward function enables LLMs to better capture complex semantic relationships while producing more balanced predictions.

Recent advances in fine-tuning suggest that incorporating explicit reasoning signals and reward optimization can significantly improve performance on semantic tasks. Approaches based on Chain-of-Thought have demonstrated gains in reasoning quality and accuracy [Liu et al. 2023, Wei et al. 2022], while reinforcement learning methods, such as Group Relative Policy Optimization (GRPO), refine outputs through relative reward signals, promoting better alignment and robustness [Shao et al. 2024, Guo et al. 2025]. Experiments show that, although SFT with Chain-of-Thought already improves over standard SFT, incorporating GRPO further enhances performance, with the proposed distance-aware reward achieving the best overall results. Our contributions are:

1. We introduce a reasoning-based formulation for theme–researcher affinity classification in the SNPG, requiring the model to produce both an affinity prediction and a structured justification grounded in semantic evidence;
2. We propose TORNADO, a two-stage fine-tuning framework for LLMs that combines reasoning learning with reward optimization for affinity classification;
3. We design task-specific reward functions, including a distance-aware reward for ordinal affinity labels, and show that the proposed strategy outperforms zero-shot, few-shot, conventional supervised fine-tuning, and GRPO with standard rewards.

2. Related Work

LLMs have demonstrated competitive performance in text classification, natural language inference, and semantic similarity analysis, benefiting from robust contextual representations acquired through large-scale pre-training. However, studies have also highlighted limitations such as hallucinations, prompt sensitivity, and inconsistent outputs, motivating the development of more robust specialization and fine-tuning strategies [Al Faraby et al. 2024]. Transformer architectures dominate classification tasks, with fine-tuning emerging as the primary strategy and generally outperforming prompt-engineering approaches [Sakai and Lam 2026]. Recent advances have explored alternative optimization schemes, including explicit reasoning mechanisms, which have shown improvements in complex semantic tasks [Mustafa et al. 2026].

In the context of affinity classification between strategic themes and researcher profiles, recent studies have focused on computational efficiency and reducing dependence on proprietary LLMs. Studies addressed a simplified binary setting (high versus low affinity) using BERT embeddings. [Miranda et al. 2025] combined MPNet and MiniLM embeddings with decision thresholds, achieving a f_1-macro of 54% but reporting many false negatives in the class high. [Utino and Gôlo 2025] evaluated lexical baselines,

similarity thresholding on BERT embeddings, and lightweight classifiers on pooled embeddings, obtaining a f_1-macro of 77% with Granite + k-NN. However, these approaches depend on model selection and remain limited by embedding aggregation strategies.

In contrast, [Sakiyama et al. 2025] combined QLoRA with a balanced loss function on Qwen3-14B through conventional fine-tuning, achieving an f_1-macro of 61% (three classes) and improving performance on minority classes, although at a high computational and energy cost. [Gôlo and Utino 2025] formulated the task as a one-class learning problem, training an independent model for each strategic theme using only high-affinity examples. This approach achieved a competitive f_1-macro of 63% but required theme-specific fine-tuning and, in some cases, was outperformed by a similarity method.

Overall, the analyzed studies reveal a recurring trade-off between computational cost, predictive performance, and robustness, as well as persistent challenges related to model sensitivity and the difficulty of capturing higher levels of semantic affinity. Although recent studies have explored the specialization of large language models through fine-tuning, most approaches still rely predominantly on conventional supervised fine-tuning. This highlights the need to investigate strategies that explicitly incorporate reasoning supervision and reward-based optimization, which are particularly suitable for affinity classification tasks, where predictions should be both accurate and explainable.

Recent GRPO-based approaches have shown that reward-guided optimization can improve model alignment and task performance beyond conventional fine-tuning [Shao et al. 2024]. The next section presents the proposed approach, which aims to balance efficiency and predictive performance through model specialization, structured reasoning, and task-specific reward mechanisms.

3. Affinity Classification with Large Language Models via Two-stage Reasoning and Reward Optimization

To address the problem of theme-researcher affinity classification in the Brazilian National Graduate System, this work proposes TORNADO, a **T**w**O**-stage **R**easo**N**ing and rew**A**r**D** **O**ptimization method for LLMs that integrates reasoning supervision and reward-based optimization [Liu et al. 2023, Chang et al. 2024, Shao et al. 2024, Guo et al. 2025]. It comprises: (i) a supervised pre-fine-tuning phase to learn structured reasoning and affinity prediction, and (ii) a reward-based fine-tuning stage using Group Relative Policy Optimization (GRPO) to refine both prediction accuracy and reasoning quality. Figure 1 illustrates the proposed framework.

To support this process, we leverage a dataset that includes affinity labels between profiles, strategic themes (*Low*, *Medium*, and *High*), and natural-language justifications supporting the assigned label. These justifications are incorporated as explicit reasoning signals during training. The overall pipeline follows a multi-stage process, from dataset preprocessing to reward-based optimization. In the pre-fine-tuning stage, the model is trained to generate structured outputs composed of two components: (i) a reasoning explanation justifying the affinity prediction; and (ii) the final classification label. Given an input q, consisting of a strategic theme and a researcher profile, the model produces an output sequence o containing both the reasoning and the prediction. To enforce this structure, special tokens are used to delimit the reasoning and answer components (e.g., `<think>` and `<answer>`). We adopt the standard supervised fine-tuning objective,

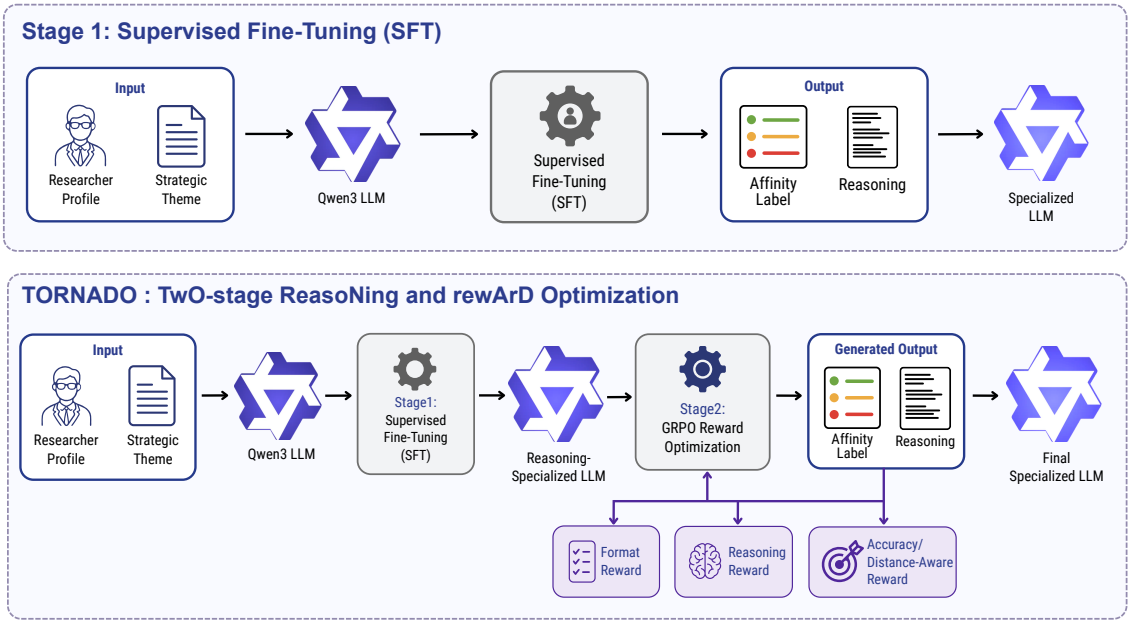


Figure 1. TORNADO method for theme-researcher affinity classification using LLMs, reasoning-based supervised fine-tuning, and reward optimization.

where the model is trained to maximize the likelihood of generating the correct sequence:

$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log \pi_\theta(o_t \mid q, o_{<t}), \tag{1}$$

where T is the number of generated tokens, q is the input prompt, and π_θ is the model parameterized by θ. This stage enables the model to learn not only the classification task but also the underlying reasoning patterns present in the dataset. After pre-fine-tuning, the model is further optimized using Group Relative Policy Optimization (GRPO) [Shao et al. 2024, Guo et al. 2025]. This stage aims to improve the model by rewarding outputs that are both accurate and semantically aligned with the expected reasoning. At each training iteration, we generate G candidate responses for each input q. Each response o_i contains a reasoning component and a predicted affinity label.

For each generated response, we compute rewards based on three complementary criteria: format compliance, semantic alignment of reasoning, and classification accuracy. We formalize these metrics by first defining the notation common to all reward functions. Let V denote the set of valid classification labels (e.g., $V = \{\text{High}, \text{Medium}, \text{Low}\}$). For a given instance, we denote the ground-truth label as $y_i \in V$ and the associated main justification as a set of words J_i, obtained strictly as the set of words with four or more characters extracted from the original justification without prior filtering. Given each o_i, we extract the predicted affinity label $\hat{y}_i$ and the set of main generated words $\hat{J}_i$ from its reasoning component.

The Format Reward (fr_i) penalizes the model when it does not comply with the XML structure specified in the prompt. Let $\mathbb{I}(o_i)$ be an indicator function that returns 1 if, and only if, the output o_i strictly contains the delimiting tags `<think>...</think>`

and <answer>...</answer>, and 0 otherwise. The Reasoning Reward (rr_i) evaluates the lexical overlap between the generated and original justifications, encouraging the learning of key elements that support the decision. It is computed proportionally to the size of the ground-truth keyword set J_i. The Format and Reasoning rewards are defined in Eqs. (2) and (3), respectively. Although we refer to rr_i as a reasoning reward, it is a lexical indicator of reasoning quality based on keyword overlap rather than a semantic evaluation of the generated explanation. The underlying assumption is that explanations sharing more relevant terms with the reference justification are more likely to contain the key elements supporting the affinity decision. We adopt this strategy instead of embedding-based similarity measures or LLM-as-a-judge approaches to avoid the additional computational cost of reward computation and training.

$$fr_i = \begin{cases} 1 & \text{if } \mathbb{I}(o_i) = 1 \text{ and } \hat{y}_i \in V, \\ 0 & \text{otherwise.} \end{cases} \quad (2) \qquad rr_i = \begin{cases} \frac{|J_i \cap \hat{J}_i|}{2|J_i|} & \text{if } J_i \neq \varnothing, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The Accuracy Reward adopts two formulations: Standard and Distance. The standard accuracy reward assigns $as_i = 2$ if $\hat{y}_i = y_i$ and $as_i = 0$ otherwise. In the Distance configuration, an ordinal error-based penalty is introduced. For this purpose, we define a mapping function $M : V \to \{0, 1, 2\}$, such that $M(\text{Low}) = 0$, $M(\text{Medium}) = 1$, and $M(\text{High}) = 2$. The distance between the prediction and the ground-truth label is given by $d(\hat{y}_i, y_i) = |M(\hat{y}_i) - M(y_i)|$. Consequently, the distance-weighted reward (ad_i) is:

$$ad_i = \begin{cases} 0 & \text{if } \hat{y}_i \notin V \\ 2.0 & \text{if } d(\hat{y}_i, y_i) = 0 \\ 0.5 & \text{if } d(\hat{y}_i, y_i) = 1 \\ -1 & \text{if } d(\hat{y}_i, y_i) = 2 \end{cases} \quad (4)$$

Let $ar_i \in \{as_i, ad_i\}$ denote the accuracy reward corresponding to the training configuration. For each response, the individual rewards are normalized independently across the G generated candidates to compute their relative advantages:

$$Ar_i = \frac{ar_i - \text{mean}(ar_1, \ldots, ar_G)}{\text{std}(ar_1, \ldots, ar_G)} \quad (5) \qquad Fr_i = \frac{fr_i - \text{mean}(fr_1, \ldots, fr_G)}{\text{std}(fr_1, \ldots, fr_G)} \quad (6)$$

$$Rr_i = \frac{rr_i - \text{mean}(rr_1, \ldots, rr_G)}{\text{std}(rr_1, \ldots, rr_G)} \quad (7)$$

The final advantage is defined as the sum of these normalized components $A_i = Ar_i + Fr_i + Rr_i$. After defining the reward functions, the objective of fine-tuning with GRPO is to maximize the advantage A_i, while ensuring that the model remains close to the reference policy through the use of the Kullback-Leibler (KL) divergence between the reference model (trained during the pre-fine-tuning stage) and the current model [Shao et al. 2024]. Consequently, the loss function is defined as follows:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\Big[\min\left(r_i A_i,\ \text{clip}(r_i, 1-\epsilon, 1+\epsilon)A_i\right) - \beta\, \mathrm{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Big], \qquad r_i = \frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}. \tag{8}$$

where clip $(\cdot, 1-\epsilon, 1+\epsilon)$ ensures that updates do not deviate excessively from the reference policy by constraining the policy ratio within the specified bounds. Specifically, q denotes the input prompt—comprising the system prompt and the theme and profile information0 in our case; π_θ represents the current trained model; $\pi_{\theta_{\text{old}}}$ refers to the model from the previous iteration; π_{ref} corresponds to the model trained during the pre-fine-tuning stage; β and ϵ are hyperparameters; and $\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ is defined by [Shao et al. 2024]:

$$\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - \log \frac{\pi_{\text{ref}}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - 1 \tag{9}$$

The central hypothesis of this work is that affinity classification is inherently a reasoning-driven task. Therefore, incorporating reasoning supervision and reward functions that capture semantic alignment and the ordinal structure of labels can improve predictive performance and model alignment. By combining reasoning-based supervision with distance-aware reinforcement learning, the proposed method promotes more accurate, robust, and explainable affinity predictions.

4. Experimental Evaluation

This section presents the experimental setup, including the dataset, model configurations, training strategies, evaluation metrics, and results. The source code is publicly available[2].

4.1. Dataset

We used the dataset released for the LeanDL-HPC Workshop challenge [Matsubura et al. 2025], which aims to foster efficient deep learning methods for High-Performance Computing (HPC) environments while considering predictive performance, memory consumption, and execution time. As a case study, the challenge focuses on mapping researcher profiles to strategic themes defined by the Brazilian Federative Units, thereby encouraging the development of solutions that produce accurate classifications and justify their decisions. The dataset contains 42,046 instances, each associating a strategic theme (represented by an identifier, name, keywords, and Federative Unit) with an academic profile (represented by an identifier, title, abstract, keywords, and field of study), along with an affinity label (*High*, *Medium*, or *Low*) and a textual justification supporting the assigned classification.

The dataset contains missing values and conflicting classifications for the same theme-output pair. To address these issues, we performed a preprocessing step considering two settings: (i) ternary (Low, Medium, and High) and (ii) binary (Low and High). In

[2]https://github.com/augustodalaldev/reasoning-reward-researcher-llm-mapping

both cases, we removed instances with missing data or simultaneously opposing classifications. Also, we applied specific procedures to each approach. In the multiclass setting, we kept records with a single classification unchanged. We consolidated records with multiple identical classifications into a single entry, arbitrarily preserving the justification from the first record. For observations with distinct but non-opposite classifications, we applied a predefined priority order — High (1), Low (2), Medium (3) — retaining the highest-priority classification and its corresponding justification. This hierarchical procedure aims to reduce a potential conservatism bias of the annotator model, which tends to concentrate classifications in the intermediate (Medium) category. In the binary approach, we filtered the dataset resulting from the previous step by removing all records labeled as Medium. Mapping these instances to the remaining classes could introduce inconsistencies between the label and the textual justification, compromising their use in this pipeline. Table 1 presents the distribution of instances across all preprocessing stages.

Table 1. Distribution of instances by affinity level throughout preprocessing.

Affinity	Original	Three-Class	Two-Class
Low	20183	15583	15583
Medium	12950	3421	-
High	8893	5665	5665

4.2. Experimental Settings

We truncated abstracts to 800 tokens using the LLM tokenizer. Exploratory analysis of full prompt lengths (prior to truncation) revealed a highly skewed distribution: the median (p50) was 1,356 tokens, 95% of the data (p95) was up to 1,944 tokens, and extreme outliers reached up to 6,020 tokens. This truncation ensured compliance with the model's context limit (2,560 tokens) and avoided implicit truncation. We generated training, validation, and test sets using a stratified split by the theme's Federative Unit in proportions of 70%, 10%, and 20%, respectively. This strategy ensures geographic representativeness across the partitions, taking into account the specificities of strategic themes.

We used the Qwen3-14B LLM, specifically its quantized version unsloth/Qwen3-14B-bnb-4bit, due to its favorable computational efficiency, multilingual capability (essential for Portuguese data), and strong reasoning ability, which is important for tasks requiring textual justification. For both classification approaches, we defined the following evaluation settings. The first two were considered baselines: the first uses a zero-shot approach, and the second employs few-shot prompting to assess the impact of in-context examples. In the remaining settings, we applied QLoRA-based fine-tuning: (i) Supervised Fine-Tuning (SFT) for direct task adaptation; (ii) SFT with Chain-of-Thought (CoT) to leverage explicit reasoning; and (iii) our proposed TORNADO approach with Group Relative Policy Optimization (GRPO) to optimize performance based on relative preferences.

We defined system and user prompts to instruct the model to classify the affinity between a strategic theme and an academic output, taking into account the specificities of each approach. In the few-shot setting, we added representative examples of theme-output pairs with their corresponding classifications, covering all possible affinity levels (three examples for the three-class case and two for the binary case).

For the SFT-CoT and GRPO approaches, we appended instructions requiring the model to include its classification and the reasoning used to support its choice within XML tags defined for this purpose (<think>...</think> and <answer>...</answer>, respectively). Once the prompt structure was defined, we established a general processing pipeline consisting of two main stages: fine-tuning and inference. In the fine-tuning stage, we executed four distinct training configurations: SFT, SFT-CoT, GRPO Standard (TORNADO-Standard), and GRPO Distance (TORNADO-Distance). While the first two methods were applied directly to the baseline, both GRPO variants were initialized with weights obtained via SFT-CoT. This choice avoids a cold start, ensuring that the model already captures task-specific nuances and the required output structure, while also promoting structured reasoning through the CoT strategy.

For parameter-efficient fine-tuning, we employed Rank-Stabilized LoRA (RS-LoRA) with rank 16 and alpha 32, targeting all linear modules of the network, resulting in approximately 0.42% of trainable parameters. The SFT phase was trained for 3 epochs using an 8-bit AdamW optimizer, a batch size of 64, and a cosine learning-rate scheduler peaking at 5e-4, with sequence packing enabled. For the GRPO, we lowered the batch size to 32 and, to prevent policy collapse, reduced the learning rate to 5e-6 for 1 epoch. The group relative advantages were calculated over 4 sampled generations ($G = 4$) with a rollout temperature of 0.7, while the KL divergence penalty (β) was set to 0.001. All final evaluations were conducted using greedy decoding (temperature 0.0).

We evaluated model performance across all combinations of classification approaches and experimental scenarios using standard metrics such as precision, recall, f_1-macro and accuracy. We ran the experiments on a Debian 12 machine with an Intel Xeon Gold 5520+ CPU, 503 GiB RAM, and an NVIDIA H100 GPU (96 GB). Throughout all training and inference phases, we systematically recorded hardware and environmental metrics, specifically peak VRAM usage, and RAM consumption. During GRPO inference, the model occasionally omitted the final label, necessitating the extraction of predictions from the reasoning tags. A small fraction of the cases (0.32%–1.13%) still did not have valid predictions, so a default "Low" label affinity was applied.

4.3. Results and Discussion

Tables 2 and 3 present the results of the experimental pipeline for the binary and ternary approaches, respectively, in the task of classifying the affinity between academic researcher profiles and strategic themes. The evaluated methods include zero-shot, few-shot, Supervised Fine-Tuning (SFT), SFT with Chain-of-Thought (SFT-CoT), and TORNADO with GRPO under the standard (TORNADO-Standard) and distance-based (TORNADO-Distance) reward configurations. The tables report macro precision, recall, F_1, and accuracy, with the best and second-best results highlighted in bold and underlined.

In the binary setting, which presents lower task complexity, even simpler approaches such as zero-shot and few-shot achieve reasonable performance (with f_1 and accuracy of approximately 0.62). SFT yields substantial gains in accuracy (0.785), albeit with an imbalance between precision (0.860) and recall (0.609), indicating a bias toward one of the classes. This behavior is mitigated by the introduction of Chain-of-Thought (SFT-CoT), which significantly improves all evaluation metrics (f_1 = 0.874 and accuracy = 0.902), demonstrating that explicit reasoning leads to more balanced predictions.

Table 2. Evaluated metrics from binary approach pipeline.

	precision	recall	f_1-macro	Accuracy
Zero-shot	0.700	0.732	0.614	0.617
Few-shot	0.703	0.737	0.622	0.625
SFT	0.860	0.609	0.616	0.785
SFT - CoT	**0.884**	0.865	<u>0.874</u>	<u>0.902</u>
TORNADO - Standard	<u>0.872</u>	**0.887**	**0.879**	**0.903**
TORNADO - Distance	0.850	<u>0.866</u>	0.857	0.885

During the reward optimization stage, TORNADO-Standard achieves the best overall performance ($F_1 = 0.879$ and accuracy = 0.903), slightly outperforming SFT-CoT. This result indicates that a reward function based solely on exact correctness is sufficient to guide the model toward improved solutions. In contrast, TORNADO-Distance provides no advantage in this setting, exhibiting a slight decrease across all metrics. This behavior is expected because in binary problems the notion of distance is less relevant, since all errors carry the same weight. Therefore, introducing differentiated penalties does not contribute to the learning process.

In the ternary setting (Table 3), task complexity increases substantially, which is reflected in the limited performance of the baseline methods. The zero-shot model achieves an f_1-macro of 0.2664 and an accuracy of 0.2882, indicating poor generalization without adaptation. Despite moderate precision (≈ 0.48), the low recall highlights the model's difficulty in correctly recovering the different classes.

Table 3. Evaluated metrics from ternary approach pipeline.

	precision	recall	f_1-macro	Accuracy
Zero-shot	0.4849	0.4201	0.2664	0.2882
Few-shot	0.4816	0.4087	0.2485	0.2740
SFT	<u>0.6207</u>	0.5039	0.5251	0.7294
SFT - CoT	0.6184	0.6159	<u>0.6151</u>	0.7567
TORNADO - Standard	0.6066	<u>0.6193</u>	0.6001	**0.7777**
TORNADO - Distance	**0.6461**	**0.6485**	**0.6467**	<u>0.7746</u>

Few-shot prompting does not improve performance and can even degrade it, suggesting that LLMs struggle with the additional context introduced by examples, likely due to limited capacity and difficulty filtering relevant information. SFT substantially improves performance, particularly accuracy (0.7294), indicating that supervised training helps the model learn discriminative patterns. However, the f_1-macro (0.5251) still reflects class imbalance. SFT-CoT further improves performance, especially on f_1-macro (0.6151), reinforcing the conclusion that the affinity task is reasoning-dependent. By making justifications explicit, the model produces more balanced predictions across classes.

In the GRPO optimization stage, TORNADO-Standard improves accuracy (0.7777) but does not yield proportional gains in f_1. This suggests that a reward function based solely on exact correctness is insufficient to capture the problem's ordinal structure.

In contrast, TORNADO-Distance achieves the best overall performance, with consistent improvements across all metrics (f_1 = 0.6467, precision = 0.6461, and recall = 0.6485). This result demonstrates that incorporating distance-based penalties enables the model to better handle errors by penalizing misclassifications between distant classes more severely while favoring predictions closer to the correct label.

A point concerns the behavior of the f_1-macro in the ternary setting, which may appear discrepant with respect to precision and recall in some cases. This occurs because the f_1 is computed as the unweighted average of the per-class f_1, making it highly sensitive to variations in performance across classes. As a result, even when aggregate metrics are high, weaker performance in specific classes can significantly reduce the overall value.

The confusion matrices shown in Figure 2 explain the quantitative results. Zero-shot and few-shot baselines produce unbalanced predictions, concentrating outputs on a few classes and frequently confusing *Low* with *High* or *Medium*. Consequently, some classes achieve high precision or recall individually but not both simultaneously, resulting in poor f_1-macro despite relatively high precision and recall. In contrast, fine-tuned models produce more balanced confusion matrices. Standard SFT improves the recognition of the majority *Low* class, but still struggles with *High* and especially *Medium*, which remains the most ambiguous class. Adding reasoning supervision through SFT-CoT reduces this imbalance and improves the diagonal distribution, suggesting that explicit reasoning helps the model distinguish affinity levels more consistently.

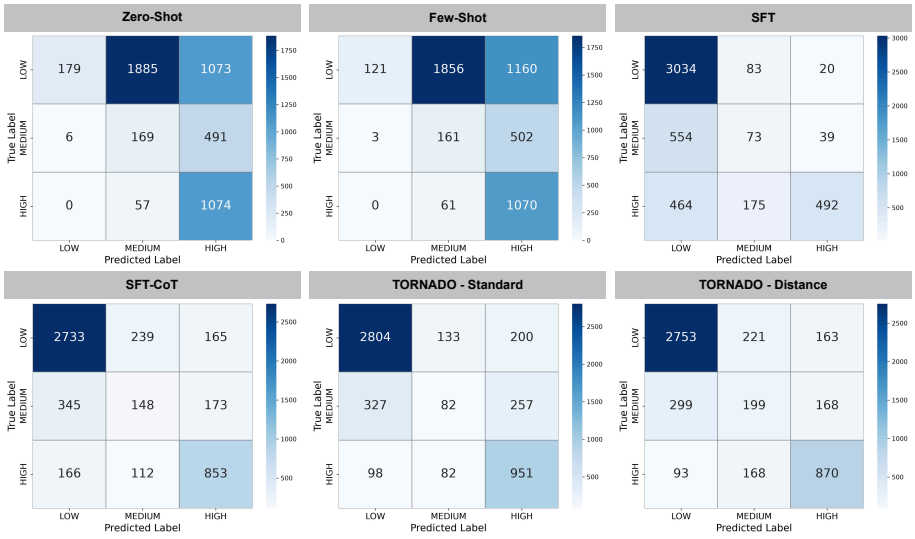


Figure 2. Confusion matrices for all methods in the ternary classification setting.

The TORNADO proposals further improve this behavior through reward-based optimization. While TORNADO-Standard achieves the highest accuracy, its confusion matrix still exhibits a considerable number of severe errors between distant classes. TORNADO-Distance, on the other hand, better captures the ordinal structure of the task by concentrating most mistakes between adjacent labels, such as *High–Medium* and *Medium–Low*, rather than between *High* and *Low*. This is important because affinity labels are not independent nominal categories, but ordered levels of semantic proximity between a researcher and a theme. Thus, adjacent-class errors may still preserve part of the underlying affinity structure, whereas distant-class errors indicate a more severe semantic

mismatch. This behavior is precisely the objective of the proposed distance-aware reward function and explains why TORNADO-Distance achieves the best f_1-macro, precision, and recall, despite a slightly lower accuracy than TORNADO-Standard. It is worth noting that TORNADO outperforms the binary and ternary approaches previously reported in the LeanDL-HPC Workshop (Section 2). Although direct comparisons are limited by differences in experimental protocols, prior results provide a useful reference for the task. Reported f_1-macro range from 0.54 to 0.77 in binary settings and around 0.61 in ternary, highlighting the challenges of modeling researcher-theme affinity.

Another relevant aspect concerns the limitations of the proposed reasoning reward. The improvements obtained by SFT-CoT indicate that exposing the model to structured justifications helps it produce more balanced predictions, especially in the ternary setting, where the Medium class is semantically ambiguous. However, the reasoning reward used in this work is based on lexical overlap with the reference justification rather than on a deeper semantic assessment of the explanation's quality. This choice reduces computational cost during GRPO training, which is consistent with the task's efficiency-oriented setting, but it may also favor explanations that reproduce relevant keywords without fully capturing the causal relationship between the profile and the theme.

5. Conclusion and Future Work

This work proposed TORNADO, a two-stage fine-tuning strategy for Large Language Models that integrates reasoning supervision with reward-based optimization. The approach combines supervised fine-tuning with reasoning (Chain-of-Thought) and a reinforcement learning stage using Group Relative Policy Optimization, enhanced with distance-based rewards designed to capture ordinal relationships between classes. The results show that reasoning supervision improves LLM performance and produces more balanced predictions. Distance-aware rewards enhance performance in settings with ordinal labels by penalizing errors according to their severity. In contrast, rewards based only on exact correctness fail to capture this structure. TORNADO improves results, particularly in the multi-class formulation of theme-researcher affinity classification.

For future work, several directions can be explored. Richer reward formulations could be investigated, incorporating semantic similarity measures or learned evaluators to better assess reasoning quality beyond lexical overlap. Alternative policy optimization strategies may further improve training stability and efficiency. Extending the approach to other structured prediction tasks, including ordinal, hierarchical, or multi-label classification problems, would help validate its generality. Additionally, integrating retrieval mechanisms or external knowledge sources may further enhance reasoning capabilities. Finally, exploring scalability to larger models and datasets can provide further insights into the trade-offs between model size, reasoning supervision, and reward design.

References

Al Faraby, S., Romadhony, A., and Adiwijaya (2024). Analysis of llms for educational question classification and generation. *Computers and Education: Artificial Intelligence*.

CAPES (2023). Plano nacional de pós-graduação – PNPG 2024–2028: versão preliminar para consulta pública. Technical report, Ministério da Educação / CAPES, Brasília.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Gôlo, M. P. S. and Utino, M. Y. R. (2025). One-class lightweight interpretable filtering for academic profiles and strategic themes affinity. In *International Symposium on Computer Architecture and High Performance Computing Workshops*, pages 147–154.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2023). Gpt understands, too. *AI Open*.

Matsubura, E. T., Borin, E., Cirne, W., Cáceres, E. N., Marcacini, R., Nogueira, B., Silva, D. F., de Andrade Silva, J., Turine, M. A. S., de Souza, M. C., et al., editors (2025). *Workshop on Lightweight and Efficient Deep Learning in HPC Environments*.

Miranda, A. C. D., Utino, M. Y. R., Gôlo, M. P. S., Dos Santos, M. A. A., and de Souza, M. C. (2025). Reducing costs in large-scale classification: A hybrid bert–llm strategy. In *Int. Symp. on Computer Architecture and High Performance Computing Workshops*.

Mustafa, A., Naseem, U., and Rahimi Azghadi, M. (2026). Can reasoning llms enhance clinical document classification? *Health and Technology*, 16.

Sakai, H. and Lam, S. S. (2026). Large language models for health care text classification: Systematic review. *JMIR AI*, 5:e79202.

Sakiyama, K. M., Fujimoto, M. L., Santin, R. V., and Rezende, S. O. (2025). Leandl hpc challenge 2025: Applying large-scale model adaptation techniques. In *Int. Symp. on Computer Architecture and High Performance Computing Workshops*, pages 139–146.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., et al. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Turine, M. A. S., Souto Maior, A. P., and Marcacini, R. M. (2025). National agenda for the training of higher-level personnel. Technical Report 1, CAPES – Coordination for the Improvement of Higher Education Personnel, Ministry of Education, Brazil.

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. Technical report, United Nations. Accessed: Apr. 20, 2026.

Utino, M. Y. R. and Gôlo, M. P. S. (2025). Efficient filtering with bert embeddings for researcher–topic affinity prediction in hpc pipelines. In *International Symposium on Computer Architecture and High Performance Computing Workshops*, pages 155–162.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.