ISSN 2669-2023 (print) / ISSN 2538-8959 (on-line) / **https://doi.org/10.47459/cndcgs.2026.55**

Cognitive Warfare and Language Models: Detecting Pro-Kremlin Narratives in the Czech Online Information Space Through Comparative Ideological Fine-Tuning

Matej RYBÁR[1]*,

[1]*Institute of Intelligence Studies, University of Defence, Address: Kounicova 65, 662 10 Brno, Czech Republic*

*Correspondence: *matej.rybar@unob.cz*

Abstract

This paper examines whether adapting a multilingual large language model (LLM) to clearly pro-Kremlin or pro-Western texts can improve the automatic detection of pro-Kremlin narratives in Czech-language content. The study compares three variants of the same model: a base multilingual model, a pro-Kremlin adapter trained on Russian-language texts with pro-Kremlin framing, and a pro-Western adapter trained on Western sources that respond to the same topics from an opposing perspective. All models are fine-tuned using a parameter-efficient method and evaluated on a small Czech corpus covering five key narrative types. The pro-Kremlin adapter shows the strongest ability to distinguish between texts with pro-Kremlin framing and neutral texts, while the pro-Western adapter brings only a modest improvement over the base model. These findings suggest that exposing an LLM to ideologically aligned training data can make it more sensitive to that ideology, with potential applications for supporting the detection of hostile information operations in smaller-language NATO member states.

KEY WORDS: *cognitive warfare, pro-Kremlin narratives, disinformation, Czech Republic, large language models, ideological fine-tuning, LoRA, QLoRA, narrative detection, information operations*.

Citation: Rybár M. Cognitive Warfare and Language Models: Detecting Pro-Kremlin Narratives in the Czech Online Information Space Through Comparative Ideological Fine-Tuning. In Proceedings of the Challenges to National Defence in Contemporary Geopolitical Situation, Brno, Czech Republic, 7-10 September 2026. ISSN 2538-8959, https://doi.org/10.47459/cndcgs.2026.55

1. Introduction

1.1. The Cognitive Warfare Context

Cognitive warfare — defined by NATO ACT as "activities conducted in synchronization with other Instruments of Power, to affect attitudes and behavior by influencing, protecting, or disrupting individual and group cognition to gain advantage over an adversary" — has become a priority concern for Euro-Atlantic defence planners [1]. Unlike traditional information operations, cognitive warfare is distinguished by its targeting of the cognitive layer of human decision-making: it attacks not just beliefs but the processes by which beliefs are formed, eroded, and revised [1]. Modern adversaries, notably Russia, exploit digital platforms and AI-enabled disruptive technologies to carry out such operations rapidly and at scale [1,13,14].

The Russian information environment has undergone a dramatic escalation since February 2022. Russia's state budget for mass media reportedly increased by 433% following the full-scale invasion of Ukraine, and the reach of Kremlin-aligned social media accounts increased substantially across Europe [13,14]. NATO and EU institutions have flagged AI-enabled disinformation and information operations as a growing strategic concern, and the World Economic Forum's 2025 Global Risks Report ranks misinformation and disinformation as the single most critical global risk [13,14]. Recent analysis by the NATO Strategic Communications Centre of Excellence (NATO StratCom COE), drawing on over 3.6 million Telegram posts and 500+ hours of Russian television, characterizes the Kremlin's propaganda architecture as a collage-like, multi-channel system in which each element — official state media, Telegram channels, television — contributes complementary dynamics of institutional legitimacy, emotional amplification, and distributed flexibility [13,14].

Peer-review under responsibility of General Jonas Žemaitis Miltitary Academy of Lithuania and University of Defence, Czech Republic

1.2. The Czech Information Space

The Czech Republic occupies a particularly salient position in this threat landscape. As an EU and NATO member providing substantial military and humanitarian support to Ukraine, Czechia has been consistently targeted by pro-Kremlin influence operations [2]. A multi-institute report by AMO, the Adapt Institute, and the EAST Center identifies a dense ecosystem of domestic and foreign actors — including media outlets such as Sputnik CZ, Aeronet, and Parlamentní listy — that serve as vectors for Kremlin-aligned narratives [2]. Research by Cvrček and Fidler at the Institute of the Czech National Corpus characterizes the discursive practice of such outlets as "parasitic": they borrow legitimacy from real news events while systematically weaving anti-West, pro-Kremlin, and anti-Ukrainian interpretative frames [2].

CEDMO's quarterly tracking data from Q1 2024 shows that the most widely spread disinformation narrative in the Czech Republic reached up to 45% of the population, with individual false claims believed by 11–58% of respondents [2]. The Russian disinformation campaign intensified further in the lead-up to the 2025 Czech parliamentary elections, which were heavily targeted by coordinated influence operations. These dynamics underscore the operational urgency of scalable, automated detection methods capable of flagging pro-Kremlin framing in Czech-language content [3-7].

Five narrative clusters dominate the pro-Kremlin Czech information space [2, 4]:

1. NATO as aggressor: depicting NATO as the provocateur of the Ukraine conflict.
2. EU as corrupt: framing EU institutions as corrupt, ineffective, or hostile to national interests.
3. Ukraine as a Western puppet: denying Ukraine's sovereign agency.
4. Sanctions self-harm: arguing that Western sanctions damage European economies more than Russia.
5. Betrayal of elites: portraying Czech political leadership as traitors serving foreign interests.

1.3. Research Gap and Contribution

Automated detection of pro-Kremlin disinformation has made significant progress using supervised transformer-based classifiers and, more recently, LLMs [3-7]. However, most existing work focuses on high-resource language settings (English, German, French) or trains models on binary disinformation labels without incorporating the ideological dimension of the training data [3,4,7,8]. The critical question of whether a model that has internalized a particular ideological framing becomes more sensitive to detecting that same framing in out-of-domain text remains underexplored [10-12].

This paper makes four contributions:

1. It operationalizes the concept of ideological fine-turning — using ideologically labeled corpora to adapt a base LLM — as a detection strategy rather than purely a bias characterization exercise.
2. It introduces a comparative adapter framework contrasting pro-Kremlin vs. pro-Western adaptation on the same base model (Qwen2-7B-Instruct).
3. It evaluates both adapters on a small Czech-language evaluation corpus covering five operationalized narrative categories.
4. It provides dual-lens evaluation combining standard classification metrics (ROC-AUC calibrated classifiers) with internal model signal analysis (PPL ΔNLL, score shift distributions) and fold-wise stability [3, 7, 20].

2. Background and Related Work

2.1. LLMs for Propaganda and Disinformation Detection

The application of large language models to propaganda and disinformation detection has developed rapidly since 2023 [3,5-7]. Sprenkamp et al. (2023) demonstrate that GPT-4 achieves performance comparable to state-of-the-art supervised models (RoBERTa-based systems) on the SemEval-2020 Task 11 propaganda detection benchmark, using a variety of prompt engineering and fine-tuning strategies [3]. However, more recent work finds that off-the-shelf LLMs without task-specific adaptation still fall short of fine-tuned baselines for fine-grained technique detection: GPT-4 yields substantially lower F1 scores than a RoBERTa-CRF system on the same dataset when asked to identify specific propaganda techniques at span level [7].

For the specific task of detecting pro-Kremlin disinformation, Kramer, Golovchenko, and Hjorth (2025) provide a landmark evaluation, demonstrating that both open-weight and closed LLMs can accurately detect pro-Kremlin tweets about the MH17 downing, outperforming both research assistants and prior supervised models — and at substantially lower cost per label [5]. Their work, however, focuses on a well-defined single-event binary classification problem and does not examine whether the LLM's performance varies with the ideological alignment of its training or adaptation data [5].

The EUvsDisinfo dataset (Leite et al., 2024), the largest multilingual resource for pro-Kremlin disinformation detection with coverage across 42 languages, demonstrates that multilingual models can be trained to distinguish pro-Kremlin from trustworthy content, and importantly reveals language-specific topical patterns that vary significantly across national

contexts [4]. Czech is among the lower-resource languages in that dataset, reinforcing the challenge of building effective Czech-language detectors with standard supervised approaches [4,8].

For Czech specifically, the Masaryk NLP group has explored using LLMs (including GPT-4-class models) to generate fine-grained span-level annotations of manipulative techniques in Czech news articles, producing annotations that "significantly enriched" an existing Czech Propaganda dataset and demonstrated that contemporary LLMs can produce useful technique-level annotations with reasonable consistency [6].

2.2. Ideological Fine-Tuning and Political Bias

A parallel body of work examines ideological fine-tuning not for detection but for characterization of LLM bias. PoliTune (Agiza et al., 2024) is the most directly relevant precursor: using Low-Rank Adaptation, it systematically biases Mistral-7B-v0.2 and Llama3-8B toward specific economic and political ideologies by curating fine-tuning data from ideologically filtered social media corpora [10]. PoliTune demonstrates that PEFT-based ideological adaptation produces measurable, directional ideological shifts without requiring full model retraining — a key insight that motivates the methodology of the present study [10].

Recent work on probing political ideology in LLMs finds that ideological directions are encoded in relatively linear subspaces of the model's representation space, and that interventions along these directions generalize across downstream political reasoning tasks including bias detection [12]. This provides a theoretical grounding for the hypothesis that fine-tuning on ideologically aligned text will shift the model's internal representations in ways that are detectable by downstream probing and classification tasks [10,12].

Olejnik (2025) raises the dual-use dimension explicitly: the same parameter-efficient fine-tuning infrastructure that enables detection research also enables adversarial influence operations, including the construction of AI propaganda factories running on commodity hardware [13,14]. This dual-use reality reinforces the strategic importance of understanding and characterizing ideological fine-tuning for defensive as well as offensive applications [10,13,14]. NATO has explicitly flagged AI-enabled influence operations as a core concern for Alliance security [13,14].

2.3. Parameter-Efficient Fine-Tuning: LoRA and QLoRA

The technical infrastructure for ideological adaptation is provided by parameter-efficient fine-tuning (PEFT) methods. Low-Rank Adaptation (LoRA; Hu et al., 2022) injects trainable rank decomposition matrices into each Transformer layer while freezing the pre-trained weights, reducing the number of trainable parameters by up to 10,000× relative to full fine-tuning while matching or exceeding full fine-tuning quality on standard NLP benchmarks [16]. QLoRA (Dettmers et al., 2023) extends LoRA with 4-bit NormalFloat (NF4) quantization, double quantization of quantization constants, and paged optimisers, making it feasible to fine-tune a 65B-parameter model on a single 48GB GPU without measurable performance degradation relative to 16-bit full fine-tuning [17].

Empirical comparisons of PEFT methods for text classification confirm LoRA's strong performance profile. The PEFT-Bench study (2026) finds that LoRA achieves the highest performance on most NLP classification tasks among a range of PEFT approaches [18]. Nwaiwu (2025) shows in a rigorous low-resource classification comparison that LoRA achieves the best F1 scores among LoRA, IA^3, and ReFT methods, reaching 0.909 F1 on Amazon Reviews with far fewer trainable parameters than full fine-tuning [19]. These benchmarks justify the choice of QLoRA as the adaptation backbone for the present study [16-19].

2.4. Cross-Lingual Transfer for Disinformation Detection

A critical methodological challenge in the present study is the cross-lingual nature of the setting: training adapters primarily on Russian and English texts, then evaluating on Czech. Cross-lingual transfer for misinformation detection has been demonstrated to be feasible using multilingual pre-trained models (mBERT, XLM-R, mDeBERTa), with English functioning as an effective source language for transfer to lower-resource languages [8]. However, transfer quality degrades with the linguistic and topical distance between training and evaluation domains [8,9].

The choice of Qwen2-7B-Instruct as the base model directly addresses this challenge. The Qwen2 series (Yang et al., 2024) demonstrates robust multilingual capabilities across approximately 30 languages, explicitly including Russian among its evaluated languages [15]. This makes Qwen2-7B-Instruct better suited than English-centric models to the present task, where the PK adapter is trained primarily on Russian-language corpora [8,15].

3. Data and Corpora

3.1. Training Corpora

The study employs two training corpora, constructed to provide ideologically contrasting fine-tuning signal. Fig. 1 summarizes their sizes and the distribution of the five narrative categories across the two training corpora and the Czech evaluation set:

Pro-Kremlin Corpus (n = 10,110 texts): Sourced from Russian-language media and digital platforms known to carry pro-Kremlin framing. The corpus covers all five target narrative categories, with a distribution peaking at the NATO as aggressor narrative (2,475 texts), followed by Ukraine as a Western puppet (2,429), EU as corrupt (1,956), sanctions self-harm (1,803), and betrayal of elites (1,447).

Pro-Western Corpus (n = 8,949 texts): Sourced from pro-Western English-language media responding to the same topics from an opposing perspective. Each text is assigned to exactly one of the five operational narrative categories. The distribution is NATO as aggressor (1,885), EU as corrupt (1,386), Ukraine as a Western puppet (2,235), sanctions self-harm (1,691), and betrayal of elites (1,752). Narrative coverage is thus structurally parallel across the two ideological corpora, with higher absolute volumes and denser framing in the pro-Kremlin corpus.

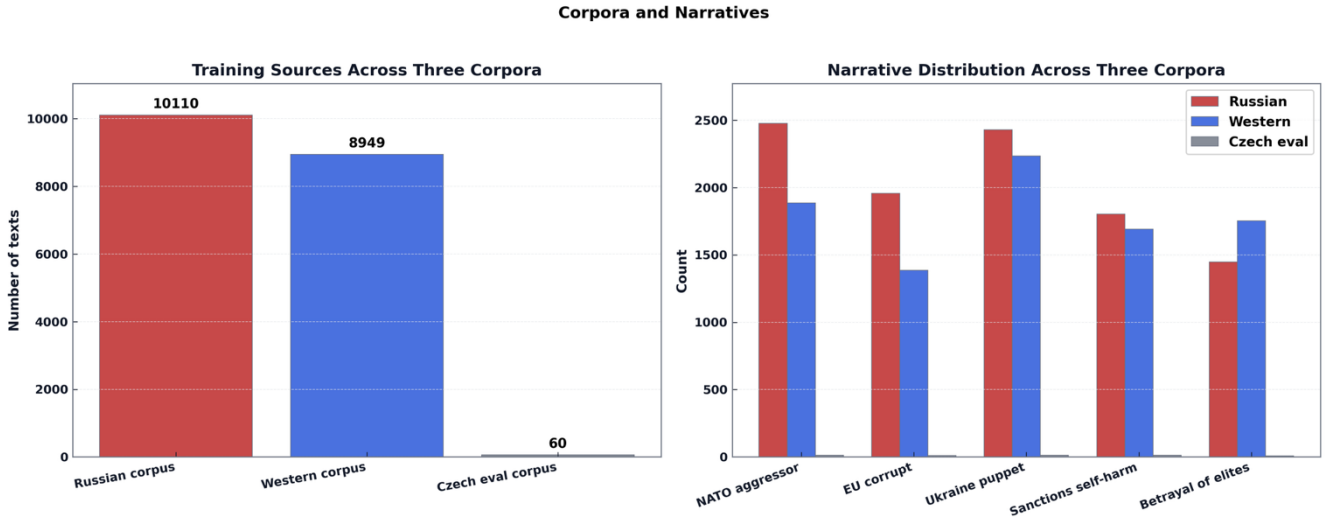

Fig. 1. Sizes of Russian, Western, and Czech corpora and distribution of five pro-Kremlin narratives across them.

This asymmetry in corpus size and narrative density reflects the operational reality of the information environment: pro-Kremlin content is more prolific across all five categories. The implication for fine-tuning is that the PK adapter trains on a stronger and more consistent ideological signal than the PW adapter, a factor that must be taken into account when interpreting detection performance differences. This imbalance is not corrected during training; instead, it is treated as an intentional feature that mirrors the asymmetric information environment rather than a controlled lab condition.

3.2. Czech Evaluation Corpus

The Czech evaluation corpus consists of 60 texts drawn from Czech-language sources and manually annotated for pro-Kremlin narrative content. The texts are distributed across the five narrative categories as follows: NATO as aggressor (14), EU as corrupt (11), Ukraine as a Western puppet (14), sanctions self-harm (12), and betrayal of elites (9). Within each category, the set deliberately balances aligned texts (with overt pro-Kremlin framing) and neutral texts, so that the overall corpus supports binary detection of pro-Kremlin framing.

The small size of this corpus (n = 60) reflects its purpose as a proof-of-concept evaluation set assembled under limited annotation and computation resources, rather than a comprehensive Czech-language benchmark, and this has important implications for the statistical interpretation of the results discussed in Section 5.

3.3. Labeling and Category Assignment

In both ideological training corpora, each text is assigned to exactly one dominant narrative category (NATO as aggressor, EU as corrupt, Ukraine as a Western puppet, sanctions self-harm, betrayal of elites) based on its main topic and framing. For the pro-Kremlin corpus, labels originate from the source datasets, which already distinguish articles by their primary pro-Kremlin narrative; these labels are adopted and mapped directly to the five operational categories. For the pro-Western corpus, it is likewise relied on existing topic-level labels from the underlying collection of Western responses to pro-Kremlin claims on the same themes, again mapping them to the five categories at the topic level rather than by stance. No automatic clustering or re-labeling is performed.

The Czech evaluation corpus is labeled independently. Each text is manually annotated by the author for (a) the presence or absence of pro-Kremlin framing (aligned vs. neutral) and (b) its dominant narrative category, if present. In the quantitative evaluation, only the binary aligned/neutral label are used and aggregated across narrative categories.

4. Methodology

4.1. Base Model

The base model is Qwen2-7B-Instruct (Yang et al., 2024), a 7-billion-parameter instruction-tuned dense transformer from the Qwen2 family [15]. Qwen2-7B offers strong multilingual capabilities across approximately 30 languages, with explicit multilingual benchmarks including Russian (ruMMLU), making it well-suited to training on Russian-language pro-Kremlin corpora and evaluating on Czech text [8,15]. As a 7B-parameter model, it is feasible to fine-tune and run on a single GPU, which matches the compute resources available for this study and makes it a realistic option for deployments outside large cloud environments [15].

4.2. Fine-Tuning with QLoRA

Both ideological adapters are trained using QLoRA (Dettmers et al., 2023): the base model is quantized to 4-bit NF4 precision, and LoRA rank-decomposition adapters are injected into the query and value projection matrices of each Transformer attention layer [16,17]. This approach freezes the bulk of the model's parameters (maintaining the general linguistic knowledge encoded during pre-training) while allowing the adapter to shift the model's internal representations toward the training distribution [16,17]. The PK adapter is fine-tuned on the pro-Kremlin corpus; the PW adapter on the pro-Western corpus. The base model receives no ideological fine-tuning.

The mathematical formulation of LoRA decomposes the weight update as:

$$W' = W + \Delta W = W + BA, \tag{1}$$

where $W \in \mathbb{R}^{d \times k}$ is the frozen pre-trained weight matrix, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are the learnable low-rank matrices, and $r \ll \min(d, k)$ is the chosen rank. Only A and B are updated during training; the pre-trained weight matrix W remains unchanged [16].

4.3. Evaluation Framework

Three complementary evaluation lenses are used:

Calibrated Classifier (ROC-AUC): For each model variant, the narrative score output is used to train a logistic regression calibrated classifier on the evaluation corpus using 5-fold cross-validation. ROC-AUC is computed for each fold and overall, providing a threshold-agnostic measure of discrimination performance appropriate for the class imbalance present in the evaluation data [3,5,7,20].

PPL Probe (ΔNLL): For each text, the negative log-likelihood (NLL) — a direct measure of the model's predictive probability over the token sequence — is computed for each of the three model variants. Perplexity (PPL) is the exponentiated NLL; here it is being worked with ΔNLL because it is additive and easier to interpret. The delta ΔNLL relative to the base model ($\Delta NLL = \Delta NLL_{base} - \Delta NLL_{adapter}$) captures whether the adapter finds the text more or less linguistically expected than the base model. A positive ΔNLL indicates lower perplexity under the adapter, suggesting that the adapter's internal representation treats the text as more consistent with its training distribution. Intuitively, a positive ΔNLL means "the adapter finds this text more linguistically expected than the base model," while a negative value means the opposite [3,7].

Fold-wise Stability: 5-fold cross-validation of the calibrated classifier provides fold-level ROC-AUC scores, allowing assessment of the variance and robustness of each model's detection signal across different train/test partitions of the small evaluation corpus. Each "fold" is a run where the classifier is trained on 80% of the texts and evaluated on the remaining 20%, rotating which subset serves as the test set [3,7,20].

These three lenses capture complementary aspects of the problem. ROC-AUC of the calibrated classifier measures how well each model variant separates pro-Kremlin and neutral texts at the decision level. The ΔNLL-based PPL probe asks whether the adapter itself finds aligned texts more "expected" than the base model, revealing changes in internal language modelling behavior. Fold-wise stability then shows how robust these signals are across different splits of the small Czech evaluation set, highlighting where apparent gains may be fragile.

5. Results

5.1. Overall Detection Performance

The PK adapter achieves the highest detection performance across both evaluation methods. The calibrated classifier yields a ROC-AUC of 0.758 for the PK model, compared to 0.622 for the base model and 0.619 for the PW model. The PPL probe confirms the directional advantage: the PK adapter achieves an NLL-based AUC of 0.719, against 0.574 for the PW probe and N/A for the base model (which serves as the reference). The base model's calibrated AUC of 0.622, still only moderately above the chance level of 0.500, confirms that generic instruction-tuning without ideological adaptation provides limited discriminative ability on this task [3,5,7].

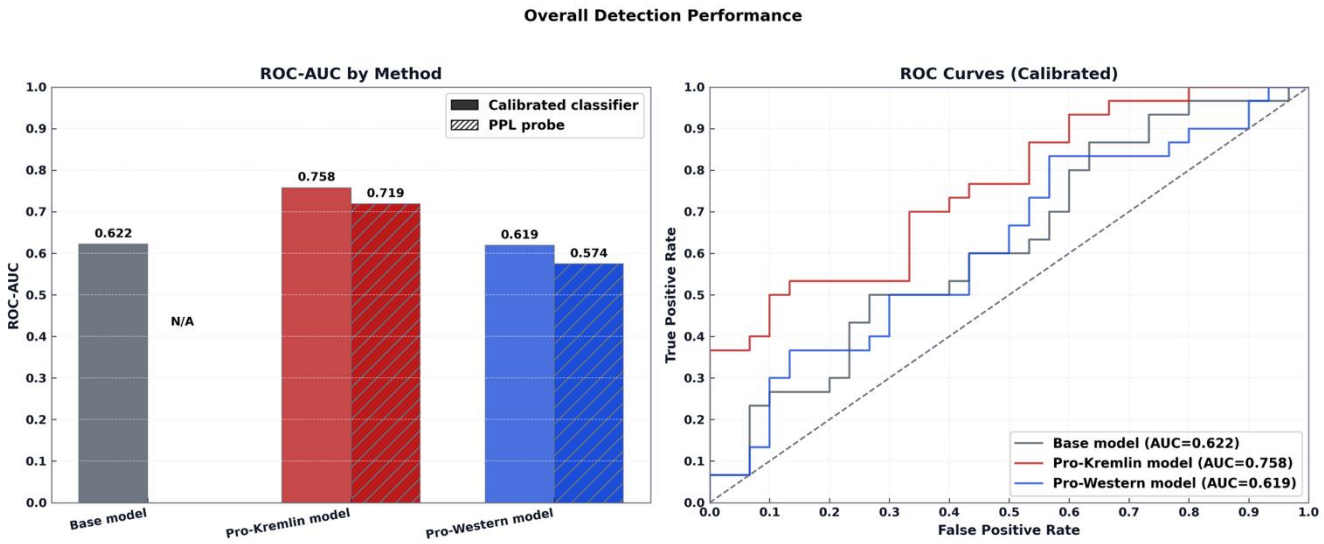

Fig. 2. ROC-AUC and ROC curves comparing base, pro-Kremlin and pro-Western models on Czech narrative detection.

The PW adapter shows a modest improvement over the base model (0.619 calibrated, 0.574 PPL probe), consistent with the hypothesis that some indirect transfer of narrative sensitivity occurs through exposure to counter-narrative Western framing. However, the effect is substantially weaker and less consistent than for the PK adapter, likely reflecting both the smaller and less narratively concentrated Western training corpus and the indirect framing relationship (counter-narratives about pro-Kremlin topics differ from the pro-Kremlin texts being detected) [4,8,10-12].

5.2. Internal Signal Analysis

The PPL ΔNLL distributions (Fig. 3, left panel) show that the PK adapter produces larger positive ΔNLL values for aligned Czech texts (mean 0.083) than for neutral texts (mean 0.028), confirming that ideological fine-tuning shifts the model's predictive distribution in a class-sensitive direction. The PW adapter exhibits a much weaker pattern: ΔNLL values are close to zero for neutral texts (mean ≈ 0.000) and only modestly positive for aligned texts (mean 0.036), consistent with its smaller detection gain [4,8].

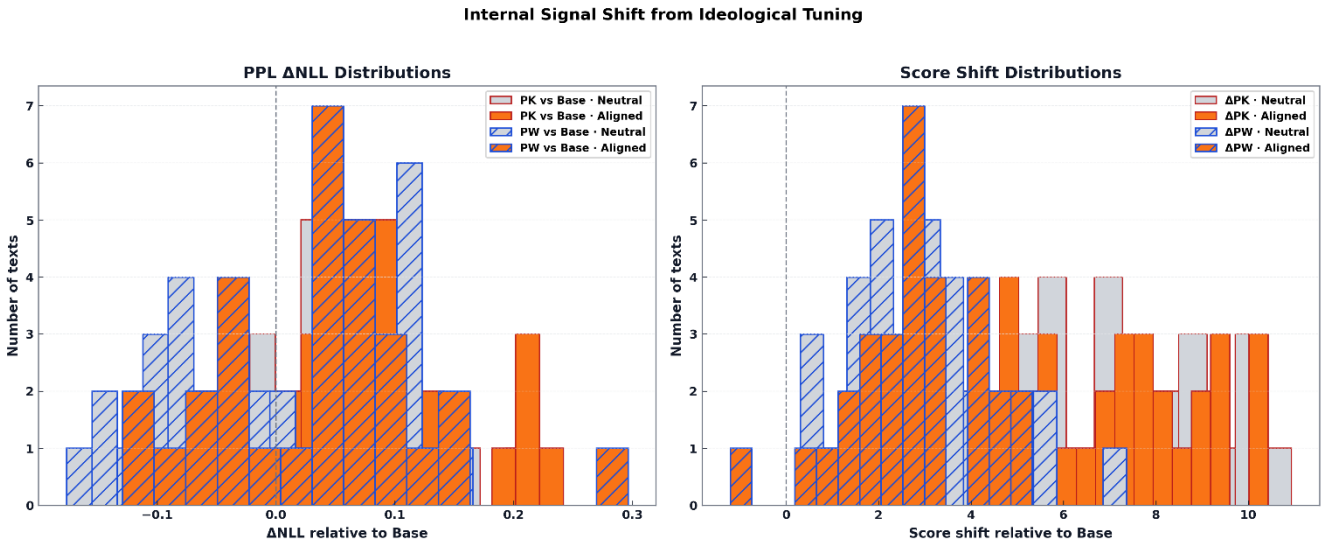

Fig. 3. ΔNLL and score shift distributions showing how Pro-Kremlin and Pro-Western tuning change internal signals.

The score-shift distributions (Fig. 3, right panel) reveal a complementary pattern. The PK score shift (ΔPK) shows higher averages for aligned texts (mean 7.33) than for neutral texts (mean 6.58), whereas the PW score shift (ΔPW) is lower overall and slightly higher for neutral texts (mean 2.83) than for aligned texts (mean 2.75). This asymmetry indicates that the PK adapter embeds more class-discriminative signal in its narrative scores, while the PW adapter primarily induces a weaker, more uniform shift [10-12,16-19].

5.3. Model Disagreement Lens

Across all 60 Czech texts, mean narrative scores are 49.8 for the base model, 56.8 for the PK adapter, and 52.6 for the PW adapter, confirming a systematic upward shift in PK scores relative to both baselines. The largest PK–PW disagreements (items A–F in Fig. 4) correspond to texts where the absolute difference between the two adapters' scores ranges from approximately 6.3 to 8.0 points, highlighting cases where the PK model's internalized pro-Kremlin framing leads it to assign substantially higher narrative scores than the PW model [3,5,7].

Fig. 4. Narrative scores per Czech text for all three models, highlighting systematic Pro-Kremlin score elevation and disagreements.

The systematic elevation of PK scores also implies a calibration challenge: the PK model's higher overall scores require threshold adjustment before deployment, since a naïve score cutoff would substantially over-flag neutral content. The calibrated classifier (logistic regression on model scores) partially addresses this, but the fold-wise instability discussed below means that calibration parameters estimated on 60 texts carry significant uncertainty [3,5,7,20].

5.4. Fold-wise Stability

The fold-wise stability analysis (Fig. 5) shows substantial variance in all three models' ROC-AUC scores across the five cross-validation folds. The PK model attains fold-level AUCs of 0.58, 0.80, 0.57, 0.81, and 0.97 (mean 0.75), remaining above chance in every fold. The base model ranges from 0.20 to 0.97 (fold AUCs 0.61, 0.20, 0.54, 0.63, 0.97; mean 0.59), and the PW model from 0.57 to 0.74 (0.67, 0.57, 0.71, 0.59, 0.74; mean 0.66). While the PK adapter maintains a consistent advantage, the wide fold-to-fold variation, especially for the base model, underscores the sensitivity of all estimates to individual texts in the small 60-item evaluation set [3,5,7].

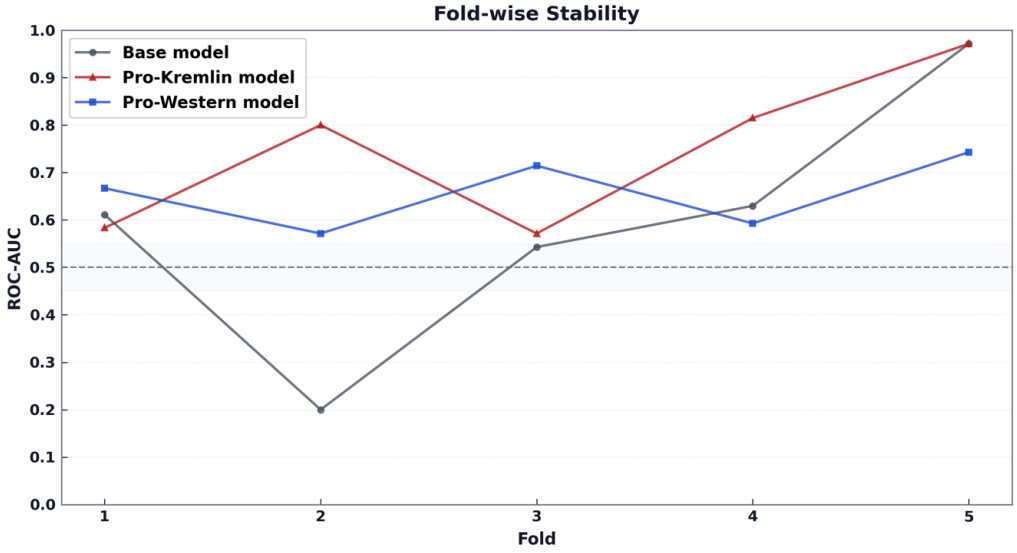

Fig. 5. Fold-wise ROC-AUC for three models, illustrating the greater stability of the Pro-Kremlin adapter.

This instability is a direct consequence of the small evaluation corpus (n = 60) and the resulting high sensitivity of fold-level train/test splits to individual text characteristics. It does not negate the PK model's directional advantage — the PK model never falls below chance while the others frequently do — but it underscores that the absolute AUC values reported in Section 5.1 carry wide confidence intervals and should not be treated as stable benchmark figures generalizable to larger or differently-distributed Czech corpora [3,5,7,20].

6. Discussion

6.1. Does Ideological Fine-Tuning Work?

The results provide consistent evidence that exposing a multilingual LLM to ideologically concentrated training data increases its sensitivity to that ideology in a held-out evaluation set. This is visible both in higher ROC-AUC for the PK adapter (0.758 vs. 0.622/0.619) and in the fact that PK ΔNLL and score shifts separate aligned from neutral texts much more strongly than PW or the base model (Fig. 2–3). This finding aligns with PoliTune's demonstration that LoRA-based fine-tuning on ideologically filtered data produces measurable, directional alignment shifts, and with the theoretical finding that ideological directions are encoded in relatively linear, transferable subspaces of LLM representations.

The mechanism is interpretable: by training on thousands of texts that employ pro-Kremlin narrative frames, the PK adapter shifts the model's representation space such that tokens and token sequences characteristic of that framing become more linguistically "expected" (lower NLL). When the model is subsequently asked to score a new Czech-language text, this representational shift manifests as elevated narrative scores for framing-consistent content — exactly the signal captured by both the PPL probe and the calibrated classifier.

6.2. Why Is the PK Adapter Stronger Than the PW Adapter?

Three factors likely contribute to the PK adapter's stronger performance relative to the PW adapter:

1. Training data volume and density: The Russian corpus (10,110 texts) is 13% larger than the Western corpus (8,949 texts) and more narratively concentrated — the five narrative categories appear more frequently and with higher intra-class consistency in Russian-language pro-Kremlin media than in Western counter-narrative sources.
2. Framing directionality: The PW corpus trains the model on counter-narratives and Western rebuttals to the same five topics. While this provides topical coverage, the framing is oppositional rather than exemplary — the model learns "here is what Western sources say in response to pro-Kremlin claims" rather than "here is how pro-Kremlin claims are actually worded." For the detection task (identifying pro-Kremlin framing), exemplary training is more informative than oppositional training.
3. Cross-lingual alignment: Qwen2's strong Russian-language capabilities mean that training on Russian corpora transfers more effectively to related Slavic framing patterns (including Czech-language content borrowing Russian narrative structures) than English-language Western corpora would.

6.3. The Dual-Use Dimension

The finding that ideological fine-tuning improves detection carries an unavoidable dual-use implication. The same QLoRA methodology used defensively in this study can be applied offensively to bias LLM outputs toward a specific ideological agenda at scale — a threat explicitly characterised as feasible with commodity hardware by Olejnik (2025). Fine-tuned LLMs are already demonstrated to produce ideologically aligned social media content that is indistinguishable from human-authored text. The NATO StratCom COE has issued specific warnings about the growing risk of AI-powered disinformation machines operating at scale with minimal human oversight.

This dual-use reality argues for treating the present methodology not as a finished product but as a research proof-of-concept that simultaneously demonstrates defensive utility and offensive risk. Responsible deployment of ideology-aware detection models requires robust evaluation on larger, independently annotated datasets before production use.

6.4. Limitations

Several limitations constrain the generalizability of these findings:

1. Small evaluation corpus (n = 60): The Czech evaluation corpus is the primary limiting factor. Collecting and manually annotating a substantially larger Czech dataset for this narrow task was beyond the scope and resources of the present experiment; therefore, the results are framed as exploratory rather than definitive. The fold-wise AUC variance visible in Fig. 5 directly reflects this limitation. Even a ROC-AUC advantage of ~0.14 over the base model, while consistent and multi-evidence, may not replicate on a larger, differently sampled Czech evaluation set [3,5,7].

2. Narrative category confounding: The five narrative categories are not mutually exclusive — texts frequently carry multiple narrative frames simultaneously — and the present evaluation aggregates across categories. Narrative-specific AUC figures would provide more operationally useful diagnostic information [2,4].
3. Cross-lingual extrapolation: The PK adapter is primarily trained on Russian-language text and evaluated on Czech. While the linguistic relatedness and documented borrowing of Russian narrative templates in Czech pro-Kremlin media supports the transfer hypothesis, the extent of this transfer has not been independently validated [2,4,8,15].
4. Threshold sensitivity: The PK model's elevated score distribution requires careful threshold calibration for operational deployment. The logistic regression calibrated classifier partially addresses this but depends on the balance of the evaluation corpus [3,5,7,20].

7. Conclusions

This paper presents evidence that ideological fine-tuning via QLoRA adapters significantly improves the detection of pro-Kremlin narratives in Czech-language text relative to an unmodified base model. A Pro-Kremlin adapter trained on 10,110 Russian-language texts achieves a calibrated ROC-AUC of 0.758 against the base model's 0.622 and a Pro-Western adapter's 0.619, with consistent supporting evidence from PPL probe analysis, score shift distributions, and model disagreement analysis. The Pro-Kremlin adapter's detection advantage is explained by a class-sensitive internal signal shift rather than a uniform score inflation, grounding the performance gain in a coherent mechanistic account.

These findings position ideological fine-tuning as a promising but immature strategy for AI-assisted cognitive warfare defence in NATO member states with lower-resource national languages. The approach is computationally efficient (QLoRA enables adaptation on a single GPU), linguistically flexible (multilingual base models handle cross-lingual narrative transfer), and interpretable (PPL probes and score distributions provide mechanistic transparency). However, production deployment requires substantially larger evaluation corpora, narrative-stratified analysis, and rigorous adversarial testing.

The dual-use nature of this research further argues for its placement within a broader defensive AI strategy that combines automated detection with human analyst oversight, adversarial robustness testing, and clear governance frameworks for the deployment of ideologically sensitive models in national security contexts — precisely the kind of coordinated approach advocated by the NATO CCDCOE and the Atlantic Council's AI warfare research agenda.

Acknowledgements. The author gratefully acknowledges the institutional support of the University of Defence and the Institute of Intelligence Studies, whose broader research environment provided valuable background and inspiration for this work. The work was supported by the Ministry of Defence of the Czech Republic through the LANDOPS project “Conduct of Land Operations” (Project No: DZRO-FVL22-LANDOPS).

References

1. **Deppe C., Schaal GS**. Cognitive warfare: a conceptual analysis of the NATO ACT cognitive warfare exploratory concept. Frontiers in Big Data, 2024. DOI: 10.3389/fdata.2024.1452129.
2. **Cvrček V., Fidler M**. From news to disinformation: unpacking a parasitic discursive practice of Czech pro-Kremlin media. Journal of Linguistics, 2024. DOI: 10.1080/00806765.2024.2317374.
3. **Sprenkamp K., et al**. Large language models for propaganda detection. arXiv preprint arXiv:2310.06422, 2023.
4. **Leite J., et al**. EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles. arXiv preprint arXiv:2406.12614, 2024.
5. **Kramer M., Golovchenko Y., Hjorth F**. Detecting pro-Kremlin disinformation using large language models. Research & Politics, 2025. DOI: 10.1177/20531680251351910.
6. **Sahitaj A., et al**. Hybrid annotation for propaganda detection: integrating LLM pre-annotation and human review. arXiv preprint arXiv:2507.18343, 2025.
7. **Jose J, Greenstadt R**. Are large language models good at detecting propaganda? arXiv preprint arXiv:2505.13706, 2025.
8. **Ozcelik, Oguzhan, et al.** Cross-lingual transfer learning for misinformation detection: Investigating performance across multiple languages. Proceedings of the 4th Conference on Language, Data and Knowledge. 2023.
9. **Alharbi S., et al**. Multimodal misinformation detection across diverse languages and modalities. University of Central Lancashire Technical Report, 2026. DOI: 10.1145/3697349.
10. **Agiza A., et al**. PoliTune: analyzing the impact of data selection and fine-tuning methods on political bias in language models. arXiv preprint arXiv:2404.08699, 2024.
11. **Vendetti V., et al**. Fine-tuning LLMs to mimic polarized social media comments. arXiv preprint arXiv:2506.14645, 2025.
12. **Zhang T**. Probing political ideology in large language models. Findings of EMNLP 2025, 2025.
13. **Olejnik L**. AI propaganda factories with language models. arXiv preprint arXiv:2508.20186, 2025.

14. **Olejnik L**. Artificial intelligence propaganda factories with language models. Applied Cybersecurity & Internet Governance. 2026. DOI: 10.60097/ACIG/215416.
15. **Yang An**, et al. Qwen3 technical report. arXiv preprint arXiv:2407.10671, 2024.
16. **Hu E. J., et al**. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021.
17. **Dettmers T., et al**. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv preprint arXiv:2305.14314, 2023.
18. **Belanec R., et al**. PEFT-Bench: A Parameter-Efficient Fine-Tuning Methods Benchmark. arXiv preprint arXiv:2511.21285, 2026.
19. **Nwaiwu C., et al**. Parameter-efficient fine-tuning for low-resource text classification: a comparative study of LoRA, IA3, and ReFT. Frontiers in Big Data, 2025. DOI: 10.3389/fdata.2025.1677331.
20. **Collot S, et. al**. Balanced Accuracy: The Right Metric for Evaluating LLM Judges -- Explained through Youden's J statistic. arXiv preprint arXiv:2512.08121, 2026.