Skoltech

MASTER'S THESIS STATUS REPORT

Leveraging Uncertainty Signals and Latent Space Structure for Control and Efficiency in LLMs

Master's Educational Program: **Data Science**

Student: **Andrey Goncharov**

Supervisor: **Alexey Zaytsev**

Moscow 2025

Contents

1 Research Problem / Questions **3**
- 1.1 Uncertainty and Reliability . . . 3
- 1.2 Multilingual Control . . . 3
- 1.3 Training Inefficiency . . . 3
- 1.4 Unifying Framework . . . 3

2 Goals and Objectives **4**
- 2.1 Primary Goal . . . 4
- 2.2 Specific Objectives . . . 4
  - 2.2.1 Objective 1: Uncertainty Characterization (COMPLETED) . . . 4
  - 2.2.2 Objective 2: Multilingual Latent-Space Control (COMPLETED) . . . 4
  - 2.2.3 Objective 3: Complexity-Aware Fine-Tuning (COMPLETED) . . . 4
  - 2.2.4 Objective 4: Integration and Thesis Completion (IN PROGRESS) . . . 4

3 Literature Review **5**
- 3.1 Uncertainty Estimation in LLMs . . . 5
- 3.2 Multilingual Representation Structure . . . 5
- 3.3 Code-Switching in Multilingual LLMs . . . 5
- 3.4 Adaptive Training Methods . . . 5

4 Methods **6**
- 4.1 Study 1: Uncertainty Characterization . . . 6
  - 4.1.1 Experimental Design . . . 6
  - 4.1.2 Uncertainty Metrics . . . 6
  - 4.1.3 Evaluation Protocol . . . 6
- 4.2 Study 2: Latent-Space Language Steering . . . 6
  - 4.2.1 Language Direction Identification . . . 6
  - 4.2.2 Steering Mechanism . . . 7
  - 4.2.3 Evaluation . . . 7
- 4.3 Study 3: Complexity-Aware Fine-Tuning . . . 7
  - 4.3.1 Pipeline Overview . . . 7
  - 4.3.2 Baselines . . . 7
  - 4.3.3 Models and Data . . . 8

5 Key Results **8**
- 5.1 Study 1: Uncertainty Estimation . . . 8
  - 5.1.1 Entropy Predicts Errors in Knowledge-Dependent Domains . . . 8
  - 5.1.2 Reasoning Requirement Modulates Uncertainty Validity . . . 8
  - 5.1.3 Model Scale Improves Calibration . . . 8
- 5.2 Study 2: Language Steering . . . 9
  - 5.2.1 Near-Perfect Language Classification from PC1 . . . 9
  - 5.2.2 Steering Reduces Code-Switching . . . 9
  - 5.2.3 Language Identity Concentrates in Final Layers . . . 9
- 5.3 Study 3: Complexity-Aware Fine-Tuning . . . 9
  - 5.3.1 Significant Accuracy Improvements . . . 9
  - 5.3.2 81% Data Efficiency Improvement . . . 9
  - 5.3.3 Entropy-Based Splitting Outperforms MASJ . . . 10

6 Thesis Structure (Planned) **10**

7 Timeline and Milestones **10**

8 List of References **11**

1 Research Problem / Questions

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse tasks, yet their deployment in production systems faces critical challenges related to **controllability** and **training efficiency**. This thesis investigates how **uncertainty signals** (token-level entropy) and **latent space structure** (hidden state geometry) can be leveraged for lightweight interventions without expensive retraining. We address three interconnected problems:

1.1 Uncertainty and Reliability

LLMs often generate responses with unjustified confidence, particularly in knowledge-intensive domains. The confidence gap between a model's certainty and actual correctness poses risks in high-stakes applications like healthcare, law, and education, where overconfidence in incorrect answers can lead to harmful decisions.

Research Question 1: *How can we reliably estimate LLM uncertainty from internal representations, and under what conditions are these estimates justified?*

1.2 Multilingual Control

Multilingual LLMs frequently exhibit unintended code-switching, generating tokens in languages different from the target despite explicit monolingual instructions. This undermines system reliability and complicates evaluation in production environments.

Research Question 2: *Can we identify and manipulate language-specific directions in LLM latent space to control code-switching without fine-tuning?*

1.3 Training Inefficiency

Current fine-tuning approaches apply uniform strategies across all data, ignoring that different question complexities require different learning approaches. This one-size-fits-all methodology leads to suboptimal performance and inefficient use of computational resources.

Research Question 3: *Can complexity-aware training strategies—informed by uncertainty metrics—improve fine-tuning efficiency while reducing data requirements?*

1.4 Unifying Framework

The central thesis is that **internal representations** of LLMs encode rich information about uncertainty, language identity, and task complexity that can be leveraged for lightweight, interpretable interventions without expensive retraining.

2 Goals and Objectives

2.1 Primary Goal

To develop a comprehensive framework for analyzing and manipulating internal representations of LLMs to improve their reliability, controllability, and training efficiency.

2.2 Specific Objectives

2.2.1 Objective 1: Uncertainty Characterization (COMPLETED)

- Develop automated pipeline for uncertainty quantification using entropy-based and model-as-judge approaches
- Validate correlation between token-level entropy and question difficulty across domains
- Analyze when uncertainty estimates are justified vs. unjustified
- **Status:** Published in arXiv:2503.01688 Sychev et al. (2025)

2.2.2 Objective 2: Multilingual Latent-Space Control (COMPLETED)

- Identify language-specific directions via PCA on parallel translations
- Develop inference-time steering method with negligible computational overhead
- Validate across multiple language pairs (English, Spanish, Russian, Chinese, Hindi)
- **Status:** Published in arXiv:2510.13849 Goncharov et al. (2025b)

2.2.3 Objective 3: Complexity-Aware Fine-Tuning (COMPLETED)

- Integrate uncertainty metrics for dataset stratification (easy/medium/hard)
- Implement differentiated training pipelines (SFT for easy, distillation for hard)
- Demonstrate efficiency gains over baseline approaches
- **Status:** Published in arXiv:2506.21220 Goncharov et al. (2025a)

2.2.4 Objective 4: Integration and Thesis Completion (IN PROGRESS)

- Synthesize findings into unified theoretical framework
- Combine techniques for end-to-end improved LLM deployment

- Complete thesis document and defense preparation
- **Status:** In progress

3 Literature Review

3.1 Uncertainty Estimation in LLMs

Uncertainty quantification for LLMs has been approached through both black-box and white-box methods. Model-as-judge approaches (Zheng et al., 2023) leverage auxiliary LLMs to evaluate response quality, while token-level probability methods (Kadavath et al., 2022) exploit the model's own confidence signals. Fadeeva et al. (2023) provided comprehensive comparisons showing entropy-based methods' effectiveness for free-form responses.

However, existing work suffers from two limitations: (1) lack of domain-specific analysis connecting uncertainty patterns to question types, and (2) failure to leverage uncertainty for downstream applications like adaptive training. Our work addresses both gaps.

3.2 Multilingual Representation Structure

Work on mBERT and XLM-R revealed that language identity concentrates in few principal components, separable from semantics in parallel texts (Conneau et al., 2020; Chi et al., 2020). Yang (2021) removed these components for language-agnostic embeddings. Wendler et al. (2024) extended this to decoder-only LLMs, revealing latent language preferences.

While prior work focused on analysis, we exploit this structure for *active generation control*, demonstrating practical code-switching mitigation.

3.3 Code-Switching in Multilingual LLMs

Code-switching has been studied for mixed-language inputs (Aguilar et al., 2020; Bali et al., 2014), with recent work identifying *unintended* switching in outputs (Ryan et al., 2024; Yoo et al., 2024). Standard mitigations require costly fine-tuning or brittle prompt engineering. Our latent-space steering offers a lightweight alternative.

3.4 Adaptive Training Methods

Curriculum learning approaches (Kim & Lee, 2024) order training examples from easy to hard, showing modest gains. The LIMA approach (Zhou et al., 2023) demonstrated that small, high-quality datasets suffice for alignment. SmallToLarge (Yang et al., 2024) uses training trajectories for data selection but requires additional model training.

Knowledge distillation (Hsieh et al., 2023) improves complex task performance but hasn't been selectively applied based on complexity. Our work bridges uncertainty estimation and adaptive training, showing that distillation is beneficial specifically for high-complexity samples.

4 Methods

4.1 Study 1: Uncertainty Characterization

4.1.1 Experimental Design

We developed an automated pipeline for evaluating uncertainty estimation on the MMLU-Pro benchmark (Wang et al., 2024) spanning 14 topics with ∼12,000 questions. Four LLMs were evaluated: Phi-4, Mistral-Small-24B, Qwen-1.5B, and Qwen-72B.

4.1.2 Uncertainty Metrics

Token-wise Entropy: For vocabulary size k and logits $\mathbf{z} = (z_1, \ldots, z_k)$, we compute:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}, \quad H = -\sum_{i=1}^{k} p_i \log p_i \tag{1}$$

Model-as-Judge (MASJ): We prompt Mistral-Large-123B to estimate (1) required education level and (2) number of reasoning steps for each question.

4.1.3 Evaluation Protocol

ROC-AUC scores quantify how well uncertainty predicts incorrect answers, stratified by:

- Subject domain (14 topics)
- Reasoning requirement (low/medium/high)
- Model architecture and scale

4.2 Study 2: Latent-Space Language Steering

4.2.1 Language Direction Identification

Given parallel corpus $\mathcal{D} = \{(s, \ell)\}_{i=1}^{N}$ with semantic content s in language ℓ, we extract hidden states $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$ from each layer. PCA identifies the first principal component as the language

direction:

$$\mathbf{v}^{(\ell)} = \arg\max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \left(\mathbf{v}^\top (\mathbf{h}_i^{(\ell)} - \bar{\mathbf{h}}^{(\ell)})\right)^2 \tag{2}$$

4.2.2 Steering Mechanism

For layers $\ell \geq \ell_{\text{crit}}$, we remove the language component:

$$\tilde{\mathbf{h}}_t^{(\ell)} = \mathbf{h}_t^{(\ell)} - s \cdot (\mathbf{h}_t^{(\ell)} \cdot \mathbf{v}^{(\ell)})\mathbf{v}^{(\ell)} \tag{3}$$

where $s \in \mathbb{R}^+$ controls intervention strength.

4.2.3 Evaluation

- **Classification accuracy:** Logistic regression on PC1 projections
- **KL divergence:** Measuring distributional shift from code-switched to steered outputs
- **Models:** Qwen2.5-1.5B, Llama-3.2-1B
- **Languages:** English, Spanish, Russian, Chinese, Hindi

4.3 Study 3: Complexity-Aware Fine-Tuning

4.3.1 Pipeline Overview

1. **Complexity estimation:** Compute answer token entropy for each training sample
2. **Data stratification:** Split dataset into easy/medium/hard terciles by entropy
3. **Differentiated training:**
    - Easy/medium: Standard SFT
    - Hard: Chain-of-thought distillation from teacher LLM

4.3.2 Baselines

- **SFT:** Uniform supervised fine-tuning on all data
- **Curriculum:** Easy $\rightarrow$ medium $\rightarrow$ hard ordering
- **Full distillation:** CoT distillation on all data

4.3.3 Models and Data

Student models: Qwen2.5-3B, Phi-4-Mini. Teacher ensemble: DeepSeek-V3, Qwen-3-235B, Llama-4-Maverick. Dataset: MMLU-Pro.

5 Key Results

5.1 Study 1: Uncertainty Estimation

5.1.1 Entropy Predicts Errors in Knowledge-Dependent Domains

Token-wise entropy achieves strong predictive performance (ROC-AUC up to 0.83 for Biology with Qwen-72B), while MASJ scores perform near-random (ROC-AUC $\approx$ 0.49).

Table 1: ROC-AUC for error prediction by domain (Qwen-72B)

Domain	**ROC-AUC**
Biology	0.83
Economics	0.80
Psychology	0.77
Physics	0.74
Mathematics	0.73
Law	0.69

5.1.2 Reasoning Requirement Modulates Uncertainty Validity

Entropy is a better predictor when no reasoning is required (ROC-AUC 0.79) compared to high-reasoning questions (ROC-AUC 0.73 for Qwen-72B). This suggests entropy primarily captures *knowledge uncertainty* rather than *reasoning difficulty*.

5.1.3 Model Scale Improves Calibration

Larger models show clearer separation between correct (low entropy) and incorrect (high entropy) predictions. Qwen-1.5B achieves the best Expected Calibration Error (ECE = 0.242) despite smaller scale, suggesting architectural factors beyond size affect calibration.

5.2 Study 2: Language Steering

5.2.1 Near-Perfect Language Classification from PC1

A single principal component enables 95-99% language classification accuracy across all tested pairs.

Table 2: Language classification accuracy (Qwen2.5-1.5B)

Language Pair	Accuracy
English - Chinese	0.98
English - Russian	0.99
English - Hindi	0.98
English - Spanish	0.95

5.2.2 Steering Reduces Code-Switching

KL divergence from original (English) distribution decreases by up to 42% after steering:

Table 3: KL divergence reduction via steering

Language Pair	Before	After
English - Chinese	8.94	5.19
English - Russian	7.78	5.43
English - Spanish	6.37	5.86

5.2.3 Language Identity Concentrates in Final Layers

Explained variance for PC1 peaks in final layers ($\sim$10% at layer 16 for Llama-3.2), with clusters emerging loosely in early layers and sharpening dramatically toward the output.

5.3 Study 3: Complexity-Aware Fine-Tuning

5.3.1 Significant Accuracy Improvements

Our pipeline achieves substantial gains over baselines:

5.3.2 81% Data Efficiency Improvement

Our pipeline matches or exceeds full distillation while using only 19% of the tokens:

- Qwen 3B: 7.98k tokens (ours) vs. 39.45k (full distillation)

Table 4: Accuracy after 20 epochs

Method	Qwen 3B	Phi4-mini
SFT (baseline)	0.39	0.51
Curriculum	0.45	0.54
Distillation (all data)	0.50	0.63
Pipeline (ours)	**0.52**	**0.64**

- Phi4-mini: 5.35k tokens (ours) vs. 30.30k (full distillation)

5.3.3 Entropy-Based Splitting Outperforms MASJ

Single-token entropy achieves ROC-AUC 0.72-0.74 for complexity estimation, while MASJ reasoning scores achieve only 0.54-0.57.

6 Thesis Structure (Planned)

1. **Introduction** – Motivation, research questions, contributions
2. **Background** – LLM architecture, uncertainty theory, representation analysis
3. **Study 1: Uncertainty Characterization** – Methods, results, implications
4. **Study 2: Language Steering** – PCA-based control, steering algorithm, validation
5. **Study 3: Complexity-Aware Training** – Pipeline design, ablations, efficiency analysis
6. **Unified Framework** – Connecting uncertainty, control, and training
7. **Discussion** – Limitations, future directions
8. **Conclusion**

7 Timeline and Milestones

Phase	Status	Completion
Phase 1: Uncertainty Characterization	Completed	Mar 2025
Phase 2: Language Control Development	Completed	Oct 2025
Phase 3: Complexity-Aware Training	Completed	Oct 2025
Phase 4: Integration & Validation	In Progress	Dec 2025
Phase 5: Thesis Writing & Defense	In Progress	Feb 2026

8 List of References

References

Sychev, P., Goncharov, A., Vyazhev, D., Khalafyan, E., & Zaytsev, A. (2025). When an LLM is Apprehensive About Its Answers – And When Its Uncertainty Is Justified. *arXiv preprint arXiv:2503.01688*.

Goncharov, A., Vyazhev, D., Sychev, P., Khalafyan, E., & Zaytsev, A. (2025). Complexity-aware fine-tuning. *arXiv preprint arXiv:2506.21220*.

Goncharov, A., Kondusov, N., & Zaytsev, A. (2025). Language steering in latent space to mitigate unintended code-switching. *arXiv preprint arXiv:2510.13849*.

Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *NeurIPS*.

Kadavath, S., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Fadeeva, E., et al. (2023). LM-polygraph: Uncertainty estimation for language models. *EMNLP System Demonstrations*.

Conneau, A., et al. (2020). Emerging cross-lingual structure in pretrained language models. *ACL*.

Chi, E.A., Hewitt, J., & Manning, C.D. (2020). Finding universal grammatical relations in multilingual BERT. *ACL*.

Wendler, C., et al. (2024). Do LLamas work in English? On the latent language of multilingual transformers. *ACL*.

Aguilar, G., Kar, S., & Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. *LREC*.

Bali, K., et al. (2014). "I am borrowing ya mixing?" An analysis of English-Hindi code mixing in Facebook. *First Workshop on Computational Approaches to Code Switching*.

Ryan, M.J., Held, W., & Yang, D. (2024). Unintended impacts of LLM alignment on global representation. *ACL*.

Yoo, H., Yang, Y., & Lee, H. (2024). Code-switching red-teaming: LLM evaluation for safety and multilingual understanding. *arXiv preprint arXiv:2406.15481*.

Yang, Z., et al. (2021). A simple and effective method to eliminate the self language bias in multilingual representations. *EMNLP*.

Kim, J., & Lee, J. (2024). Strategic data ordering: Enhancing large language model performance through curriculum learning. *arXiv preprint*.

Zhou, C., et al. (2023). LIMA: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Yang, Y., et al. (2024). SmallToLarge (S2L): Scalable data selection for fine-tuning large language models. *arXiv preprint arXiv:2403.07384*.

Hsieh, C.-Y., et al. (2023). Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. *ACL*.

Wang, Y., et al. (2024). MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *NeurIPS*.

Appendix: Publications and Resources

A. Published Papers

1. **Sychev, P., Goncharov, A., Vyazhev, D., Khalafyan, E., & Zaytsev, A.** (2025). "When an LLM is Apprehensive About Its Answers – And When Its Uncertainty Is Justified." *arXiv:2503.01688*.

    - GitHub: https://github.com/LabARSS/question-complexity-estimation

2. **Goncharov, A., Vyazhev, D., Sychev, P., Khalafyan, E., & Zaytsev, A.** (2025). "Complexity-aware fine-tuning." *arXiv:2506.21220*.

    - GitHub: https://github.com/LabARSS/complexity-aware-fine-tuning

3. **Goncharov, A., Kondusov, N., & Zaytsev, A.** (2025). "Language steering in latent space to mitigate unintended code-switching." *arXiv:2510.13849*.

    - GitHub: https://github.com/fxlrnrpt/language-steering-in-latent-space

B. Key Findings Summary

Table 5: Summary of thesis contributions

Study	Key Method	Main Finding	Metric
Uncertainty	Token-wise entropy	Predicts errors in knowledge domains	ROC-AUC 0.83
Language Control	PCA + projection	Controls code-switching	42% KL reduction
Complexity Training	Entropy stratification	Matches distillation with 19% data	0.52/0.64 accuracy

C. Datasets Released

- MMLU-Pro with token probability distributions
- Chain-of-thought responses with entropy annotations
- Parallel translation embeddings for language steering