Skoltech

MASTER'S THESIS STATUS REPORT

From Internal Representations to Output Distributions: Improving Reliability, Control, and Training Efficiency in Large Language Models

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

Student: **Andrey Goncharov**

Supervisor: **Alexey Zaytsev**

Moscow 2025

Contents

1 Research Problem / Questions **3**
- 1.1 Uncertainty Characterization . . . 3
- 1.2 Latent Space Structure . . . 3
- 1.3 Training Inefficiency . . . 4
- 1.4 Unifying Theme . . . 4

2 Goals and Objectives **4**
- 2.1 Primary Goal . . . 4
- 2.2 Specific Objectives . . . 4
  - 2.2.1 Objective 1: Output Distribution Analysis . . . 4
  - 2.2.2 Objective 2: Latent Space Structure Exploration . . . 5
  - 2.2.3 Objective 3: Uncertainty-Guided Efficient Training . . . 5
  - 2.2.4 Objective 4: Synthesis and Thesis Completion . . . 5

3 Literature Review **5**
- 3.1 Output Distributions: Uncertainty Estimation . . . 6
- 3.2 Internal Representations: Interpretable Structure . . . 6
- 3.3 Code-Switching as a Testbed . . . 6
- 3.4 Adaptive Training Methods . . . 6

4 Methods **7**
- 4.1 Study 1: Output Distribution Analysis . . . 7
  - 4.1.1 Experimental Design . . . 7
  - 4.1.2 Signals from Output Distributions . . . 7
  - 4.1.3 Evaluation Protocol . . . 7
- 4.2 Study 2: Latent Space Structure Exploration . . . 7
  - 4.2.1 Probing for Interpretable Structure . . . 7
  - 4.2.2 Steering as Validation . . . 8
  - 4.2.3 Evaluation . . . 8
- 4.3 Study 3: Uncertainty-Guided Efficient Training . . . 8
  - 4.3.1 Pipeline Overview . . . 8
  - 4.3.2 Baselines . . . 8
  - 4.3.3 Models and Data . . . 8

5 Preliminary Results **9**
- 5.1 Study 1: Output Distribution Signals . . . 9
  - 5.1.1 Output Entropy Predicts Errors in Knowledge-Dependent Domains . . . 9
  - 5.1.2 Reasoning Requirement Modulates Signal Validity . . . 9
  - 5.1.3 Model Scale Improves Calibration . . . 9
- 5.2 Study 2: Latent Space Structure . . . 9
  - 5.2.1 Language Identity is Linearly Encoded . . . 9
  - 5.2.2 Steering Validates Causal Role of Structure . . . 10
  - 5.2.3 Structure Emerges Progressively Through Layers . . . 10
- 5.3 Study 3: Uncertainty-Guided Efficient Training . . . 10
  - 5.3.1 Significant Accuracy Improvements . . . 10
  - 5.3.2 81% Data Efficiency Improvement . . . 10
  - 5.3.3 Output Entropy Outperforms MASJ for Complexity . . . 11

6 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: **unreliable outputs** (hallucinations, miscalibrated confidence), limited **controllability**, and **inefficient training**. This thesis investigates how to improve reliability, control, and training efficiency by leveraging signals from both **internal representations** (hidden state geometry) and **output distributions** (token-level entropy) at two stages:

- **Training time:** Output entropy guides data stratification for more efficient fine-tuning (Study 3).
- **Runtime:** Output entropy enables uncertainty characterization without additional inference (Study 1), while latent space structure enables generation control without retraining (Study 2)—with potential for future efficient runtime interventions.

The key insight is that the forward pass of an LLM produces signals that can be exploited to address these challenges: output distributions reflect model confidence and uncertainty, while hidden states encode high-level structure that can be manipulated for control. This thesis addresses three interconnected problems:

1.1 Uncertainty Characterization

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 characterize LLM uncertainty using token-level entropy, and under what conditions are these uncertainty signals informative?*

1.2 Latent Space Structure

LLMs encode high-level concepts in their hidden states, but the geometry of these representations remains underexplored. Understanding how ideas—such as language identity, topic, or style—are represented in latent space opens the door to direct manipulation of model behavior without retraining. This thesis presents early exploration of latent space structure, using language identity as an ideal test case: it is discrete, easily verifiable, and has immediate practical applications (e.g., mitigating unintended code-switching). Success here suggests potential for future efficient runtime interventions on other concepts.

Research Question 2: *Can we identify interpretable structure in LLM latent space, and can manipulating this structure control generation behavior?*

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 uncertainty signals guide complexity-aware training strategies to improve fine-tuning efficiency while reducing data requirements?*

1.4 Unifying Theme

The central thesis is that reliability, control, and training efficiency can be improved by leveraging two complementary signal sources from the LLM forward pass:

- **Output distributions:** Token-level entropy reflects model confidence and task complexity, enabling runtime uncertainty characterization to flag potential hallucinations (Study 1) and training-time data stratification for more effective fine-tuning (Study 3).
- **Internal representations:** Hidden states encode interpretable structure (e.g., language identity) that can be directly manipulated for runtime generation control. This thesis presents early exploration of latent space geometry (Study 2), demonstrating feasibility and opening directions for future interventions on other failure modes.

Together, these signals provide practical approaches to improving reliability, control, and training efficiency.

2 Goals and Objectives

2.1 Primary Goal

To improve LLM reliability, control, and training efficiency by leveraging signals from internal representations and output distributions: output entropy for uncertainty characterization (improving reliability) and efficient training, and latent space structure for generation control (with this thesis providing early exploration toward future interventions).

2.2 Specific Objectives

2.2.1 Objective 1: Output Distribution Analysis

- Develop automated pipeline for uncertainty quantification using entropy from output distributions
- Validate correlation between token-level entropy and question difficulty across domains

- Analyze when output entropy signals are informative vs. misleading
- **Status:** Published in arXiv:2503.01688 Sychev et al. (2025)

2.2.2 Objective 2: Latent Space Structure Exploration

- Investigate how high-level concepts are encoded in hidden state geometry
- Use language identity as a case study: identify language directions via PCA
- Develop inference-time steering to demonstrate practical control applications
- 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: Uncertainty-Guided Efficient Training

- Use output entropy for dataset stratification by complexity (easy/medium/hard)
- Implement differentiated training pipelines (SFT for easy, distillation for hard)
- Demonstrate fine-tuning efficiency gains over baseline approaches
- **Status:** Published in arXiv:2506.21220 Goncharov et al. (2025a)

2.2.4 Objective 4: Synthesis and Thesis Completion

- Articulate unified framework: improving reliability, control, and training efficiency via internal representations and output distributions
- Discuss practical implications and future directions (extending latent space exploration)
- Complete thesis document and defense preparation
- **Status:** In progress

3 Literature Review

This thesis draws on two complementary research traditions: (1) work analyzing output distributions for uncertainty estimation, and (2) work analyzing internal representations for interpretability and control.

3.1 Output Distributions: Uncertainty Estimation

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 output distributions. 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 output entropy to question types, and (2) failure to leverage these signals for downstream applications like adaptive training. Our work addresses both gaps.

3.2 Internal Representations: Interpretable Structure

Work on mBERT and XLM-R revealed that high-level concepts—particularly language identity—concentrate in few principal components of hidden states, 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 interpretable geometric structure in internal representations.

While prior work focused on analysis, we present early exploration of exploiting this structure for *active generation control*. Language identity serves as an ideal case study: it is discrete, verifiable, and has immediate applications (code-switching mitigation). This opens the door to future efficient runtime interventions on other concepts encoded in latent space.

3.3 Code-Switching as a Testbed

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. We use code-switching mitigation as a practical demonstration of latent space steering, validating that manipulating internal representations can control generation behavior.

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 output distribution analysis and adaptive training, showing that distillation is beneficial specifically for high-complexity samples as identified by output entropy.

4 Methods

4.1 Study 1: Output Distribution Analysis

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 Signals from Output Distributions

Token-wise Entropy: For vocabulary size k and output 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 Structure Exploration

4.2.1 Probing for Interpretable Structure

We investigate whether high-level concepts are encoded as linear directions in hidden state space. Language identity serves as our case study due to its discrete, verifiable nature. 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 as Validation

To validate that the identified structure is causally meaningful, we intervene on hidden states during inference. 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: Uncertainty-Guided Efficient Training

4.3.1 Pipeline Overview

1. **Complexity estimation:** Use output entropy as a signal to estimate complexity 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 Preliminary Results

5.1 Study 1: Output Distribution Signals

5.1.1 Output Entropy Predicts Errors in Knowledge-Dependent Domains

Token-wise entropy from output distributions achieves strong predictive performance (ROC-AUC up to 0.83 for Biology with Qwen-72B), while MASJ scores perform near-random (ROC-AUC ≈ 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 Signal Validity

Output 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 output distribution signals primarily capture *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: Latent Space Structure

5.2.1 Language Identity is Linearly Encoded

A single principal component of hidden states enables 95–99% language classification accuracy across all tested pairs, confirming that language identity is encoded as a linear direction in latent space.

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 Validates Causal Role of Structure

Intervening on the identified language direction causally affects generation. KL divergence from original (English) distribution decreases by up to 42% after steering, demonstrating that latent space structure can be exploited for control:

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 Structure Emerges Progressively Through 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. This reveals how the model progressively refines high-level concepts, and identifies where latent space structure becomes most amenable to intervention. The concentration in later layers also connects internal representations to output distributions, as these layers most directly influence the final logits.

5.3 Study 3: Uncertainty-Guided Efficient Training

5.3.1 Significant Accuracy Improvements

Our pipeline, using output entropy to guide data stratification, 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)
- Phi4-mini: 5.35k tokens (ours) vs. 30.30k (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**

5.3.3 Output Entropy Outperforms MASJ for Complexity

Single-token output entropy achieves ROC-AUC 0.72–0.74 for complexity estimation, while MASJ reasoning scores achieve only 0.54–0.57. This validates output distributions as a reliable signal source for guiding training strategies.

6 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	Stage	Signal Source	Main Finding	Metric
Output Distributions	Runtime	Token entropy	Predicts errors in knowledge domains	ROC-AUC 0.83
Latent Space Structure	Runtime	Hidden states	Language encoded linearly; steering controls generation	42% KL red.
Uncertainty-Guided Training	Training	Token entropy	Matches distillation with 19% data	0.52/0.64 acc.

C. Datasets Released

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