Transforming Causal LLM into MLM Encoder for Detecting Social Media Manipulation in Telegram

2025

Prepared by: Anton Bazdyrev

Introduction & Motivation

Disinformation on social media poses significant threats to public discourse and democratic processes. In the Ukrainian context, Telegram is a primary channel for news dissemination and propaganda, where rhetorical manipulation techniques can influence opinions without factual support. Accurate detection of these techniques at both the document and span levels is crucial for fact-checking,media literacy, and automated moderation.

Evaluation Metrics

Technique Classification

Macro-averaged F_1

$$F_1$$

Span Identification

Character-level F_1

Local Validation

$$\text{balance} = \frac{F_1}{2 - F_1}$$

LB probbing

Multi-label Stratified K-Fold (CLS Labels)

Threshold Optimization

01

Technique Classification

02

03

04

Span Identification

Maximizing Grid Search

$t_{\text{gs}} = \arg\max_t \ \mathrm{F}_{1\,\text{val}}(t)$

Class-Balance Regularization

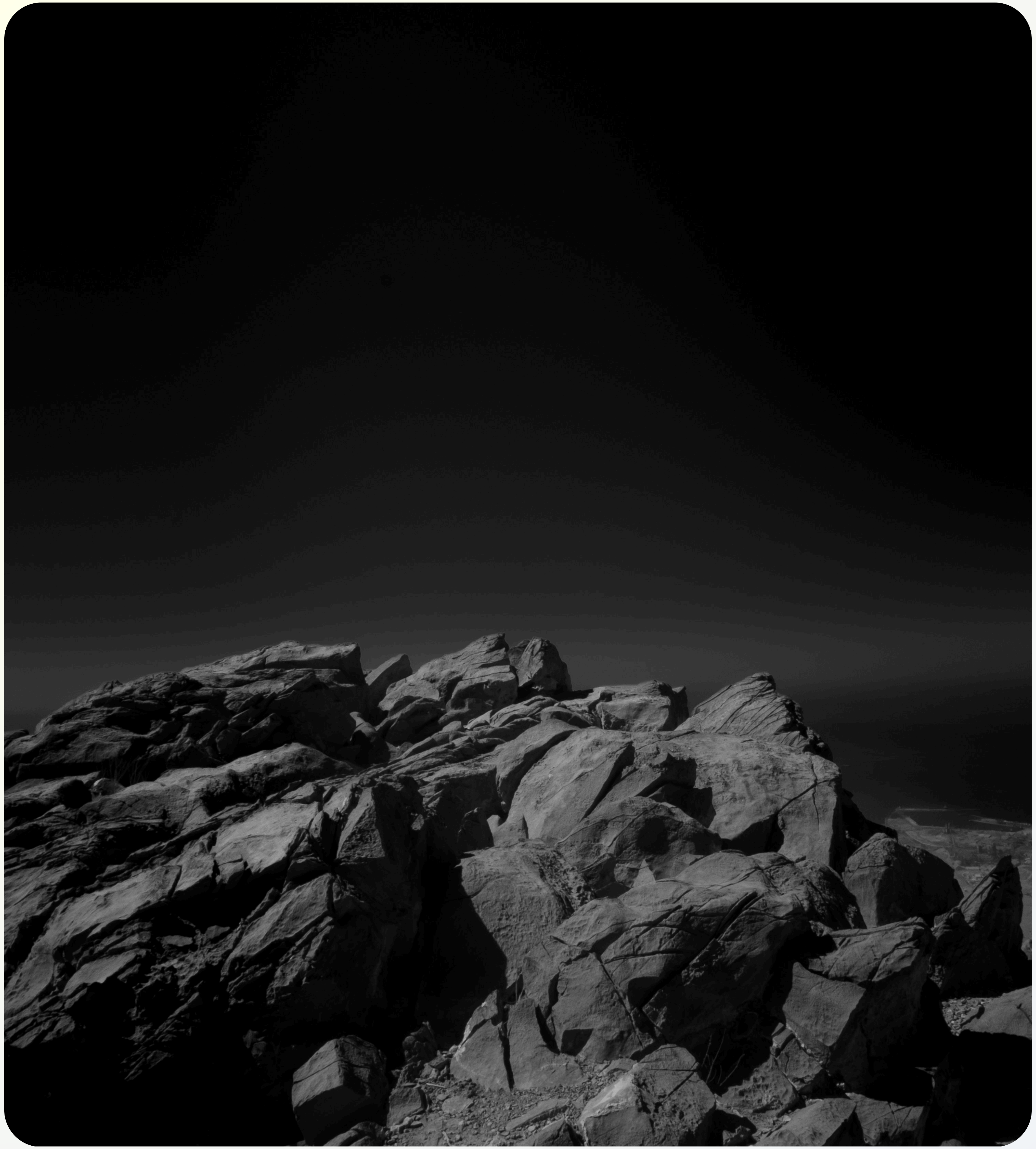
$t_{\text{cb}} = \arg\min_t \left| r(t) - r^* \right|$

Alternative Method

Thresholding method provided by Zachary C. Lipton (2014)

Hybrid Threshold

$t_{\text{final}} = \alpha\, t_{\text{gs}} + \beta\, t_{\text{cb}}$



Experimental Setup:
Technique Classification

We conducted a series of experiments with such models as **Aya-Expanse**, **LLaMA3**, and **Mistral-Large** on held-out validation data, evaluating our competiton metric.

Gemma2 consistently **outperformed** all alternatives, demonstrating superior capacity to capture nuanced patterns in the text.

Accordingly, **Gemma2-27B** was adopted as the **core architecture** for our classification pipeline.

Technique Classification

Model	Local Validation	Public LB	Private LB
Gemma2-27b (ensemble)	-	**0.474**	**0.494**
Gemma2-27b	**0.500**	0.460	0.481
Gemma2-9b	0.496	0.440	0.480
Gemma3-27b	0.483	0.439	0.468
Gemma2-27b (Lipton)	0.493	0.428	0.457
Gemma2-2b (translated)	0.413	0.375	0.370
Aya-Expanse-8b	0.419	0.389	0.414
Aya-101	0.307	-	-
LLaMA3.2-3b translated texts	0.410	0.334	0.357
Phi-4	0.412	-	-
Mistral-Large-123b	0.458	-	-

Experimental Setup:

Span Identification

- The nature of the sequence labeling task requires models to be capable of bidirectional contextual understanding. Consequently, our experiments were primarily focused on **encoder-only** architectures, including models such as **mBERT, XLM-RoBERTa, EuroBERT, mDeBERTaV3, Aya-101 (encoder)**.

- We also investigated whether large-scale architectures with robust pretraining could overcome their inherent unidirectional limitations. We experimented with **decoder-only** architectures, including **Mistral, Phi4, LLaMA3, Gemma2, Gemma3**.

Span Identification

Baselines

Model	Local Validation	Public LB	Private LB
Gemma3-27b	**0.633**	**0.615**	**0.613**
Gemma2-27b	**0.627**	**0.610**	**0.611**
LLaMA3.3-70b	0.547	-	-
LLaMA3.1-8b	0.581	0.570	0.572
Mistral-Large-123b	0.599	-	-
Aya-101 (encoder)	**0.628**	**0.611**	**0.613**
mDeBERTa-v3	**0.624**	**0.610**	**0.612**
EuroBERT-2b	0.566	-	-
mT5 (encoder)	0.572	-	-
No ML solution	0.396	0.393	0.389

Gemma Model: Exceptional Performance

On the span-identification task, our best decoder-only model, **Gemma2-27B**, achieved results which matched or slightly **outperformed** leading **encoder-only** baselines — e.g., Aya-101 and mDeBERTa-v3.

Demonstrating that even without bidirectional attention, large causal Gemma models can capture sufficient context for competitive span detection.

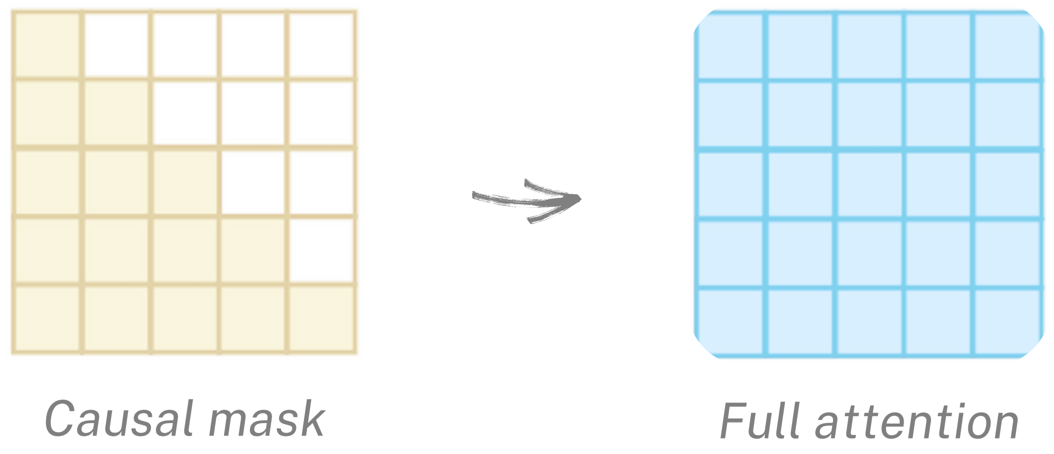
To improve boundary precision, we chose to transform **Gemma2** into a **bidirectional encoder**, enabling full left-and-right context, and then fine-tune it for span identification.

This decision was driven by the hypothesis that bidirectional representations would more reliably capture span edges.

Biderctional decoder training pipeline

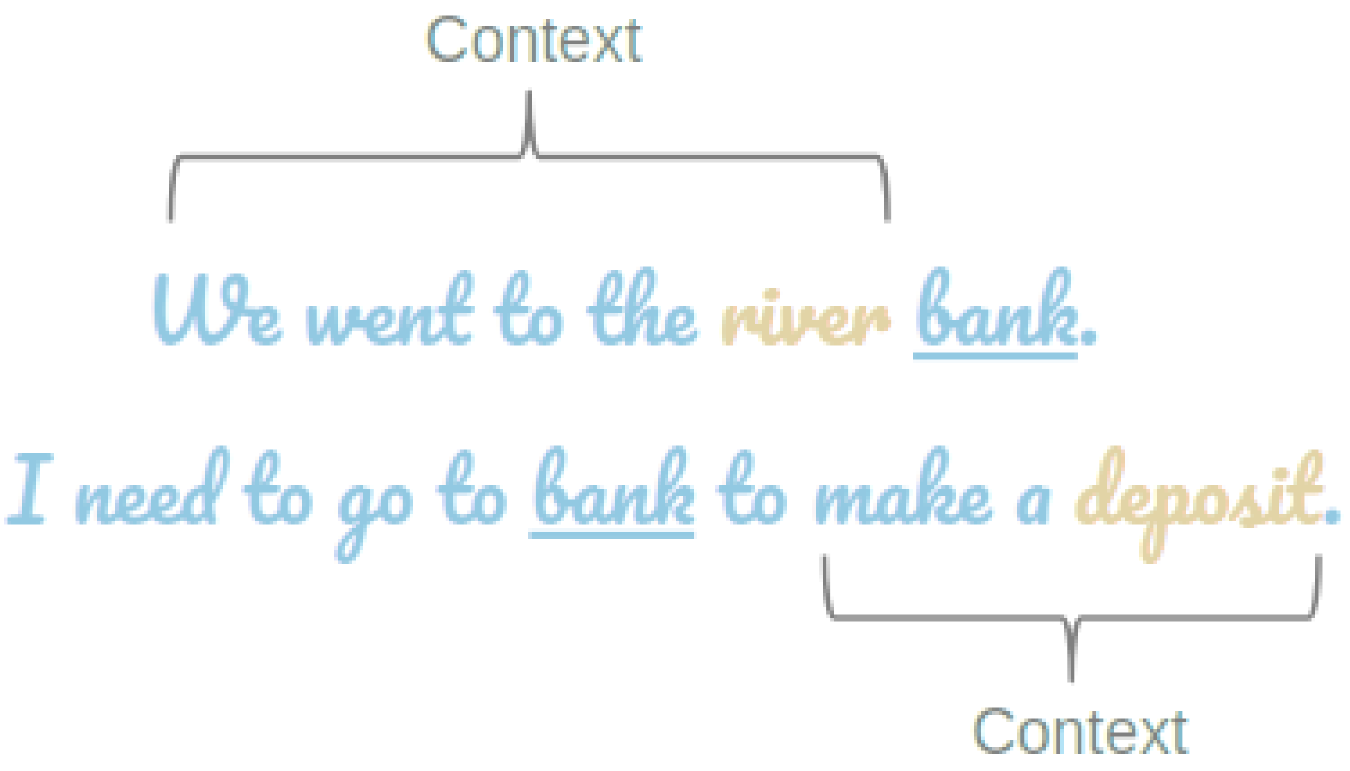
Causal Unmasking via Masked Language Modeling (MLM)

We conducted MLM pretraining on domain-related corpra to improve Gemma2's bidirectional context modeling capabilities, which resulted to what we call the **biGemma2** encoder model.



Span Identification Fine-tuning

Subsequently, we fine-tuned the model specifically for span identification, optimizing its ability to detect token-level manipulation.



Span Identification

Model	Local Validation	Public LB	Private LB
biGemma2-27b/Aya-101/mDeBERTa-v3 (ensemble)	-	**0.646**	**0.642**
biGemma2-27b (ensemble)	-	0.646	0.641
biGemma2-27b	**0.650**	0.641	0.640
biGemma2-9b	0.646	0.632	0.637
Gemma3-27b	0.633	0.615	0.613
Gemma2-27b	0.627	0.610	0.611
biLLaMA3.1-8b	0.611	0.615	0.614
LLaMA3.3-70b	0.547	-	-
LLaMA3.1-8b	0.581	0.570	0.572
LLaDA-8b	0.553	0.540	0.542
Mistral-Large-123b	0.599	-	-
Aya-101 (encoder)	0.628	0.611	0.613
mDeBERTa-v3	0.624	0.610	0.612
EuroBERT-2b	0.566	-	-
mT5	0.572	-	-
No ML solution	0.396	0.393	0.389

Alternative Approaches

01 Translation-Based Classification

Translate Ukrainian posts into English and apply LLaMA3/Gemma2 for multilabel technique classification. Despite the strong performance of these models in English, translation noise and domain mismatch degraded macro-F_1 compared to directly trained Ukrainian models.

02 Zero-Shot Classification

Use GPT-4o in zero-shot mode ($F_1 \approx 0.32$) and chain-of-thought prompting ($F_1 \approx 0.36$) to identify manipulation techniques. These relatively low scores highlight label inconsistencies and ambiguous class boundaries, suggesting potential issues with label reliability.

03 LLaDA Span Detection

Evaluate the bidirectional diffusion model LLaDA for token-level span identification. Despite its scale and novel architecture, it underperformed mDeBERTa and Gemma2 - likely due to language/domain adaptation challenges.

04 Two-Stage Positive-Only Pipeline

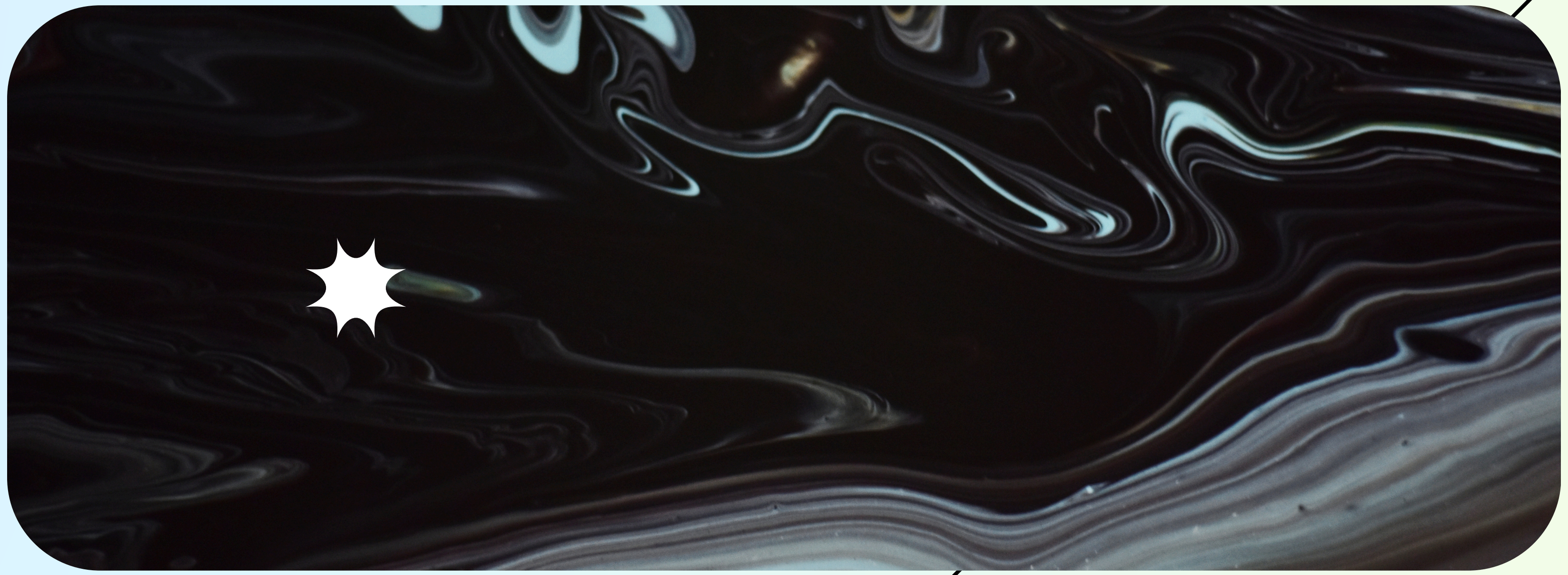
First, detect whether a post is manipulative with a binary classifier; then apply a dedicated span identifier on positives. This cut down spurious spans on clean text but introduced error propagation, yielding lower character-level F_1 than our end-to-end baseline.

05 Joint Multi-Task Learning with Auxiliary Loss

Jointly fine-tuned mDeBERTa/Gemma2 with classification and span heads via an auxiliary loss; stable training but no F_1 gains over separate models due to task interference.

Conclusions

- Gemma Competitiveness
- Threshold Optimization
- Bidirectional Pretraining
- Dual-Task SOTA Results



Thank you for your attention