

Detecting Toxic Communication in Hungarian: Theory, Corpus, and Thread-Conditioned Classification

Péter Hatvani

PPKE Doctoral School of Linguistics
hatvani.peter@hallgato.ppke.hu

Gábor Madarász

Magyar Telekom Nyrt.
Madarasz.Gabor@ext.telekom.hu

Abstract

Toxicity detection in Hungarian has been limited to message-level classification. We argue that there is a *conceptual* and a *resource* deficit. First, we introduce a three-test operational framework grounded in Culpeper’s impoliteness theory, distinguishing face-aggression (**T1**), antisocial perlocutionary effect (**T2**), and community norm violation (**T3**). T2 is already operationalised by the prior Jigsaw-scheme classifier (Hatvani and Yang, 2025); the present paper addresses the two dimensions that prior work left open: T1 and T3. Decomposing the construct achieves norm-violation agreement (T3) comparable to the prior scheme and reveals face-aggression (T1) as a harder sub-problem that binary annotation had conflated within a single label. Second, we present a silver-annotated corpus of 148,474 Hungarian conversational comments in 35,431 threads, with a 485-item gold-annotated subset, following the thread-level conversational design of Chang and Danescu-Niculescu-Mizil (2019). Third, we introduce **ModernBERT-HU-4k**, a new Hungarian encoder pretrained with a 4,096-token context window and a purpose-built tokenizer achieving 1.52 fertility on Hungarian; we fine-tune it for thread-contextual toxicity classification. This work advances the agenda of Amironesei and Díaz (2023), who called for thread-level, relational instantiations of offensive-speech modelling in non-English settings.

1 Introduction

Hungarian online discourse is almost entirely absent from sequential toxicity modelling, and the dominant computational operationalisation (the Jigsaw “likely to make you leave a discussion” criterion (Wulczyn et al., 2017)) conflates phenomena that pragmatic theory treats as distinct: direct face-aggression against an interlocutor, per-

locutionary harm to the conversational trajectory, and violation of community behavioural norms. The only publicly available Hungarian classifier (Hatvani and Yang, 2025) (huBERT-SetFit, $F_1 = 0.937$) operates message-by-message under this single-label scheme ($\kappa_{\text{free}} = 0.525$); this F_1 score is not directly comparable to the present work, as it was obtained on a different task (T2 perlocutionary harm) evaluated on that classifier’s own held-out test set. No Hungarian conversational corpus analogous to *Conversations Gone Awry* (Zhang et al., 2018) exists.

This paper addresses both a conceptual and a resource deficit. We introduce a three-test operational framework grounded in Culpeper’s impoliteness theory (Culpeper, 2011), distinguishing face-aggression (**T1**), antisocial perlocutionary effect (**T2**), and community norm violation (**T3**). T1 and T3 are the legally relevant dimensions: face-aggression maps onto conduct recognised as actionable under Hungarian and EU law (defamation, harassment, hate-speech incitement), and norm violation underpins platform compliance obligations under the EU Digital Services Act (European Parliament and Council of the European Union, 2022); T2 is a behavioural perlocutionary metric without equivalent legal standing. T2 is already operationalised by prior work (Hatvani and Yang, 2025), so the present paper targets T1 and T3. We construct a large-scale corpus of **148,474** Hungarian conversational comments in **35,431** threads, with a **485-item** human gold-annotated subset spanning five annotation layers. Finally, we introduce **ModernBERT-HU-4k**, a new Hungarian encoder pretrained with a 4,096-token context window and a purpose-built tokenizer (fertility 1.52), fine-tuned for thread-contextual toxicity classification following the conversational thread design of Chang and Danescu-Niculescu-Mizil (2019).

The thread-level design is empirically motivated: Pavlopoulos et al. (2020) show that con-

¹This paper contains examples of offensive and abusive language.

versational context shifts roughly 5% of human toxicity judgements (concentrated at the contested boundary where content alone does not determine harm) and Zhang et al. (2018) demonstrate that toxic trajectories can be predicted from a conversation’s opening moves, before any overtly harmful content appears. Decomposing the construct along T1/T3 achieves norm-violation agreement comparable to the prior single-label scheme while revealing face-aggression as the harder, context-dependent sub-problem that binary annotation had conflated. This work advances the agenda of Amironesei and Díaz (2023), who called for thread-level, relational instantiations of offensive-speech modelling in non-English settings.

Main finding. The primary contribution is the T1/T2/T3 framework: a pragmatic decomposition of a construct that Hungarian work has so far treated as one binary label. The gold corpus is the instrument that operationalises the framework, and ModernBERT-HU-4k is the demonstration that its harder dimension (T1, face-aggression) can be learned from thread context by a small encoder. The corpus and the model are evidence of the framework.

Contributions.

- A three-test (**T1/T2/T3**) operational framework grounded in Culpeper’s impoliteness theory (§2).
- A 485-item human gold corpus of Hungarian comments annotated in full thread context, with T1 and T3 binary labels (with confidence ratings) as the primary annotation layer driving all model training and evaluation, plus three secondary annotation layers (target face type, impoliteness mechanism, and Hungarian-specific form) retained in the released corpus for future work; and a 148,474-comment silver dataset generated by the best zero-shot LLM (§4), publicly released.¹
- A zero-shot benchmark of 11 LLMs on the gold corpus, establishing Gemini 3 Flash as the closest to human gold ($F_1 = 0.769$).
- **ModernBERT-HU-4k**: a new Hungarian encoder with a 4,096-token context window (tokenizer fertility 1.52), fine-tuned for thread-contextual toxicity classification and scoring within the confidence interval of the top-

ranked zero-shot LLMs (§5). Both the pre-trained encoder and the fine-tuned classifier are publicly released.²

2 Theoretical Framework

Toxicity as a multi-test construct. Recuero (2024) argues that toxicity is not a property of utterances but of *platformised discourse environments*: contexts in which platform affordances determine what content spreads and acquires social legitimacy, thereby shaping what poisons subsequent interaction. Amironesei and Díaz (2023) sharpen this further: offence is a property of *social relations*, not of linguistic tokens: the same expression is toxic or benign depending on who addresses whom, in what community, with what history.

This relational view implies that no single criterion can identify toxicity across all social configurations: a message must be assessed for direct aggression against the hearer (T1), its likely effect on the conversational trajectory (T2), and its alignment with the community’s behavioural norms (T3), since these dimensions can decouple from one another. We formalise this as a three-test operational construct:³

T1: Face-aggression. The message instantiates a conventionalised or implicational impoliteness formula in the sense of Culpeper (2011): an insult, threat, dismissal, silencer, or sarcastic mock-politeness targeting the hearer’s quality face, social-identity face, equity rights, or association rights (Spencer-Oatey’s rapport-management dimensions).

T2: Antisocial perlocutionary effect. The message produces, or in the annotation context is judged likely to produce, a disruptive perlocutionary effect: hearer exit, silencing, retaliation, or conversation derailment. This test operationalises the Jigsaw “likely to make you leave” criterion (Wulczyn et al., 2017) and is already addressed by a prior Hungarian classifier (Hatvani and Yang, 2025), which was built on exactly this definition; T2 annotation is therefore outside the scope of the present work.

²Pretrained encoder: <https://huggingface.co/magyar-nlp-szine-java/ModernBERT-HU-4k>. Fine-tuned classifier: <https://huggingface.co/magyar-nlp-szine-java/ModernBERT-HU-4k-thread-toxic>.

³Following Culpeper (2011) and standard impoliteness research, the tests are ordered from most localised (speaker-hearer dyad) to most distributed (community norm).

¹<https://huggingface.co/datasets/magyar-nlp-szine-java/silver-toxic-split>

T3: Community norm violation. The message violates the *politic behaviour* (Culpeper, 2011) expected by competent native participants in the given platform genre and community. This test is culturally and temporally relative: what is politic on kuruc.info (an extremist Hungarian online newspaper) differs from what is politic on a parenting forum.

Conjunctive and disjunctive cases. In prototypical cases (explicit slurs, direct threats) all three tests fire; the utterance is unambiguously toxic. At the periphery, the tests decouple: ritual banter among friends fires T1 but not T2 or T3 (not toxic); toxic positivity (an ostentatiously warm message that functionally silences distress) fires T2 and T3 but not T1 (it violates community norms of authentic emotional support while containing no lexically aggressive content, remaining invisible to profanity-based classifiers). Recording which tests fire, rather than a single binary label, preserves this structure; in the present work we annotate and model T1 and T3, with T2 covered by Hatvani and Yang (2025).

Hungarian-specific impoliteness vectors. Hungarian morphology provides aggression mechanisms absent in English: (i) T/V address-form shifts (*tegezés*(tu)/*magázás*(vos)) are a face-attack without lexical equivalent (Brown and Gilman, 1960); (ii) diminutive (*-kal-ke*) and hypocoristic (*-csi/-i*) suffixes whose prototypical function is affectionate become vectors of sarcastic condescension when deployed in an inappropriate register, making morphology itself the aggression carrier; (iii) euphemised taboo forms (*ba*zeg*, *wa*ze*) carry the same illocutionary force as the base forms while evading lexical profanity filters. These vectors are invisible to classifiers trained on English or generic multilingual data, even to embedding models (§5).

3 Related Work

Message-level toxicity detection. The Perspective API (Wulczyn et al., 2017) standardised the perlocutionary definition of toxicity; Pavlopoulos et al. (2020) document its context-blindness, and Hoffmann et al. (2025) show that classifiers degrade sharply across political topics, confirming that toxicity is indexically bound to its time-place-target configuration.

Sequential and conversational toxicity detection. Zhang et al. (2018) introduced the *Conversations Gone Awry* (CGA) corpus; Chang and Danescu-Niculescu-Mizil (2019) generalised this with CRAFT, a hierarchical encoder that emits per-turn derailment probabilities. Madukwe et al. (2020) survey hate-speech datasets across languages, confirming the absence of conversational sequential resources beyond English; no Hungarian sequential-derailment corpus exists.

Modular and multi-dimensional annotation. Khurana et al. (2022) demonstrate that task-specific definitions built from modular criteria outperform universal schemas; Melis et al. (2025) confirm that the definition chosen changes zero-shot LLM classification output. Ron et al. (2023) show that hate speech decomposes into independent discursive factors (contempt, abuse, incitement, prejudice, denial), anticipating the multi-test approach taken here.

Relational and critical perspectives. Deep classifiers embody a behaviourist language ideology that structurally misreads covert and implicational abuse (Castelle, 2018), motivating the inclusion of implicational face-aggression variants within T1. Cercas Curry et al. (2024) show that hate speech and offence are orthogonal: ideological categories warrant high-agreement annotation, while individual offence is inherently subjective. Cross-cultural bias in Anglophone datasets (Tonneau et al., 2024) further supports the need for a Hungarian-specific corpus.

Hungarian NLP infrastructure. HuBERT (Nemeskey, 2021) established the first competitive Hungarian BERT encoder but is limited to 512 tokens, making it unsuitable for multi-turn thread encoding without truncation. The only prior Hungarian work (Hatvani and Yang, 2025) produced a HuBERT-based SetFit (Tunstall et al., 2022) classifier. We address the context-length bottleneck with ModernBERT-HU-4k (§5), a new long-context encoder model following the ModernBERT architecture (Warner et al., 2025). There has been no Hungarian sequential toxicity benchmark created before.

4 Corpus

The corpus construction pipeline proceeds in four stages before human annotation begins: pseudonymisation, thread assembly, model-guided

pre-screening, and toxicity prevalence estimation. See 1.

4.1 Resources

Five Hungarian-language platforms were selected to cover a broad range of registers and toxicity levels (Table 6): **Reddit** (39 Hungarian subreddits, general to professional topics; collected via PRAW (The PRAW Development Team, 2024)); **YouTube** (59 curated playlists centred on polarising public figures and news channels; YouTube Data API v3); **TikTok** (daily top Hungarian-tagged videos; TikTok Research API); **Kuruc.info** (Hungary’s most prominent radical outlet, included for high-toxicity and far-right register coverage); and **Mandiner.hu** (a conservative news portal, representing a belittling subgenre of offensive discourse). Kuruc and Mandiner were scraped via headless-browser automation (Playwright; Microsoft Corporation 2024) due to the absence of public APIs.

4.2 Pseudoanonymisation

All raw data are pseudoanonymised prior to annotation following Hatvani et al. (2023): personal names, usernames, e-mail addresses, and URLs are replaced with typed placeholders (e.g. [SZEMÉLY]/[PERSON], [URL]), with surrounding Hungarian case suffixes preserved (*Jánosnak* → [PERSON]+DAT). Public figures are excluded, as their public-sphere activity carries no heightened privacy risk under EU data protection law.

4.3 Thread Assembly

Comment threads are reconstructed from the reply-to graph of each platform. A thread is defined as the connected component rooted at a top-level post, retaining temporal order.

4.4 Pre-screening and Prevalence Estimation

The full pool of assembled threads is scored with the prior HuBERT-SetFit classifier (Hatvani and Yang, 2025), which predicts whether at least one Jigsaw-category toxic message is present per thread. The scores serve two roles: estimating platform-level toxicity prevalence and providing the stratification variable for gold-annotation sampling.

Prevalence estimation. Toxicity prevalence varies substantially across platforms: Mandiner and Kuruc (both radical Hungarian outlets) flag

over 73% of comments as toxic, while TikTok reaches only 24% (full per-source breakdown in Table 6, Appendix A). Thread-level positive rates are uniformly high (65%–99%), because a thread is flagged as soon as any constituent comment is positive. Candidate threads for gold annotation were sampled in proportion to each platform’s weak-label toxicity rate, so the annotation pool preserves cross-platform prevalence differences rather than correcting for them.

4.5 Human Gold Annotation

Sampling unit. The gold unit is a single *comment* judged in context, not a thread. From each of the five sources, 100 comments are drawn, stratified on the weak toxicity label so the sample mirrors that source’s toxic/neutral split rather than balancing it. This yields 500 target comments with known source provenance. Each target comment M_n is presented together with its full reply chain $M_0, \dots, M_{n-1}$, reconstructed from the parent-of graph and shown in temporal order; the antecedent turns are read-only context and are not labelled. Because sampling is over comments, many targets are thread-initial: the mean position of a target in its chain is 0.73 (median 0), so the context manipulation in §6 is a genuine ablation only for the subset of targets that have antecedent turns.

Annotation item. An **annotation item** is one (target comment, reply chain) pair. Exactly one comment per item, the final one, receives labels; everything earlier is context. The 485 evaluation items are drawn from gold threads containing 862 comments in total (1.73 per thread on average).

Annotation scheme. Annotators apply a two-tier scheme to the target comment M_n .

Tier 0, core labels. T1 (face-aggression) and T3 (community norm violation), each a binary variable with a confidence rating on a 1 to 4 scale (1 = very uncertain, 4 = fully certain).

Tier 1, secondary labels (only when T1 or T3 fires). Target face type (Dim II), impoliteness mechanism (Dim III), pragmatic function (Dim IV), and Hungarian-specific form (Dim V). Coverage in the gold corpus falls off sharply across these fields (Dim II 47%, Dim III 41%, Dim IV 19%, Dim V 16%), which is why only T1 and T3 are carried into silver annotation and modelling.

Annotation platform and annotators. Annotation was conducted via an Argilla instance hosted on Hugging Face Spaces. Every item is annotated independently by two volunteers. An item is escalated for adjudication when the pair’s mean agreement across four fields (T1, T3, Dim II target, and Dim V Hungarian form) falls below 0.5, with binary fields scored 1 or 0 on equality and multi-label fields scored by Jaccard overlap; 129 of the 498 items met this criterion and were re-annotated by a third volunteer, who had not seen the item, in a separate -round3 dataset. For an adjudicated item the third rater’s label is final; the full label-resolution rule is given in §4.5. Six annotators participated: all were adult native Hungarian speakers aged 28–42 (gender-balanced), unpaid volunteers, five of six with a university-level background. Prior to annotation, all annotators received theory-based training covering the T1/T3 framework, the distinction between profanity and toxicity, and worked through representative examples; guidelines were embedded directly in the Argilla interface. Participation was fully anonymous: no personal or census data were linked to individual responses; only an irreversible hash is stored as the annotator identifier in the released dataset. Participants could withdraw at any time without consequence. Psychological consultation was offered to all annotators given exposure to toxic and hateful content.

Gold corpus accounting. Table 1 tracks the gold set from sample to evaluation. Of 500 sampled target comments, 498 were reconstructed with their reply chains and loaded for annotation. 497 received at least one annotation; these are the reference set for the zero-shot LLM benchmark (Table 3), which scores a model against whatever human label exists. 485 received at least two independent annotations and a resolved label on both core tests; these, and only these, are the evaluation set for inter-annotator agreement (Table 2) and for every model result (Tables 4 and 5). The 12 single-annotation items are not consensus gold and are excluded from all evaluation; they and the 1 unannotated item are retained only in the released corpus. Two annotators contributed most of the work (459 and 390 items); four others and one adjudication-round rater completed the rest, for six annotators in the main pass and a seventh in adjudication.

Inter-annotator agreement. Table 2 reports Krippendorff’s α and observed agreement on

Table 1: Gold corpus accounting. The paper uses two denominators: 497 for the zero-shot LLM benchmark (Table 3) and 485, the multiply-annotated set with resolved labels, for all agreement and model results (Tables 2, 4, 5).

Stage	Items
Target comments sampled (100 per source $\times$ 5)	500
Loaded for annotation (reply chain reconstructed)	498
≥ 1 annotation (zero-shot LLM benchmark set)	497
routed to a third rater (pair agreement < 0.5)	129
≥ 2 annotations, resolved labels (evaluation set)	485
single annotation only (released, not evaluated)	12
no annotation	1

Table 2: Inter-annotator agreement on the doubly-annotated subset ($n = 485$). Krippendorff α and observed agreement for core dimensions.

Dimension	Value	Target
T1 face-aggression (α)	.300	.65
T1 observed agreement	66.8%	
T3 norm violation (α)	.511	.60
T3 observed agreement	79.2%	

the 485 items. **T3** (community norm violation) reaches $\alpha = 0.511$ with 79.2% observed agreement, meeting the moderate-agreement threshold and approaching the pre-specified 0.60 target. **T1** (face-aggression) yields $\alpha = 0.300$ (66.8% observed), fair by conventional standards but below the 0.65 target. Compared with the prior Jigsaw-scheme corpus (Hatvani and Yang, 2025) ($\kappa_{\text{free}} = 0.525$), T3 holds its agreement under a theoretically grounded definition; the two coefficients are computed differently on different data, so the comparison is indicative only. T1 identifies a harder sub-problem absent from the prior binary scheme. Treating one annotator’s labels as predictions against the other’s yields an inter-annotator F_1 of 0.564 (T1) and 0.664 (T3) on the doubly-annotated set, establishing the human-level agreement baseline against which models are implicitly compared.

Label resolution. Each core label is resolved by the first applicable rule: the shared value where the two annotators agree (324 T1, 383 T3); the third rater’s label where they disagree and the item was adjudicated (106 T1, 80 T3); otherwise the label of the annotator with the higher stated confidence, or the earlier submission where confidence is tied (55 T1, 22 T3). Confidence is never averaged into a label and acts only as a tiebreaker; as a standalone judgement it shows near-zero agree-

ment ($\alpha \approx 0.02$ to 0.04).

Where the T1 disagreement sits. T1 disagreement (170 of 485 items, 35.1%) is not uniform. **By platform**, it falls from 55.7% on kuruc and 44.9% on mandiner to 30.9% on Reddit, 26.0% on YouTube and 17.7% on TikTok (Table 11), tracking the density of borderline implicational impoliteness rather than of overt profanity. **By mechanism**, agreement on the finer Dim III impoliteness label is lower again (53 to 80% disagreement across platforms), which is why Dim III is retained only as an exploratory layer. **By annotator**, the two primary annotators alone, on their 361 shared items, reproduce the same gap (Cohen’s $\kappa = 0.31$ for T1 versus 0.48 for T3), so the low α is not an artefact of the smaller contributors. **By confidence**, the 129 adjudicated items stay split under a third independent rater, so these are genuinely contested cases rather than annotation error, consistent with Cercas Curry et al. (2024). Conclusions about T1 modelling are therefore reported as indicative (see Limitations).

LLM benchmark. The gold corpus serves as a reference for zero-shot LLM annotation quality. All 11 models received the same zero-shot annotation prompt (reproduced with English translation in Appendix A.9), which specifies the T1 and T3 definitions from §2 in Hungarian and instructs the model to return a four-field JSON object without secondary dimensions. Exact model identifiers sufficient to reproduce the benchmark are listed in Appendix A.5. Outputs were scored against human gold using accuracy, Cohen’s κ , and F1 (Table 3).

For **T1** (face-aggression), the best single-model F1 is achieved by Gemini 3 Flash ($F_1 = 0.769$, $\kappa = 0.648$), followed by Nemotron Ultra 550B ($F_1 = 0.751$, $\kappa = 0.610$). A consistent pattern across all evaluated models is high precision but low recall: LLMs systematically underpredict aggression and label borderline implicational impoliteness as benign. The two Hungarian-specific models, PULI-LlumiX-Llama-3.1 Chat (Yang et al., 2025) ($F_1 = 0.359$) and Racka 4B (Csibi et al., 2026) ($F_1 = 0.515$), lag substantially behind larger multilingual models.

For **T3** (community norm violation), all models perform considerably better, and the ranking shifts: Gemini 3.1 Pro achieves the highest agreement ($F_1 = 0.796$, $\kappa = 0.709$), followed by Gemini 3 Flash ($F_1 = 0.769$, $\kappa = 0.662$). The consis-

Table 3: Zero-shot LLM performance versus human gold on T1 and T3 ($n = 485$). Acc: accuracy; κ : Cohen’s kappa; F1: positive-class F1. †: Hungarian-specific model. **Highlighted row**: model selected for silver annotation (best T1 F1 among all evaluated models).

Model	T1 aggr.			T3 norm		
	Acc	κ	F1	Acc	κ	F1
Claude Opus 4.8	.773	.479	.621	.843	.602	.705
Claude Sonnet 4.6	.796	.539	.675	.849	.629	.733
DeepSeek V4 Pro	.777	.491	.633	.845	.613	.717
Gemini 3 Flash	.839	.648	.769	.854	.662	.769
Gemini 3.1 Pro	.808	.576	.716	.878	.709	.796
GPT-OSS 120B	.759	.463	.633	.821	.595	.727
MiniMax M3	.821	.599	.724	.839	.583	.685
Nemotron Ultra 550B	.819	.610	.751	.814	.586	.724
Qwen 3.7 Max	.763	.447	.582	.831	.552	.655
PULI-LlumiX-Llama-3.1						
Chat†	.690	.246	.359	.818	.555	.681
Racka 4B†	.728	.363	.515	.726	.334	.527

tent T1–T3 performance gap confirms that norm violation is a more surface-accessible signal than dyadic face-aggression.

Silver corpus. Based on the LLM benchmark, Gemini 3 Flash is selected for large-scale silver annotation: it achieves the best T1 F1 among all evaluated models and competitive T3 performance, and its inference cost is favourable at scale. Only T1 and T3 labels are generated for the silver corpus; Dim II through Dim V are excluded based on insufficient benchmark agreement.

Silver selection. Silver annotation is restricted to comments detected as Hungarian with at least 0.5 confidence by lingua-py (Mistahl, 2026) and containing at least two Latin alphabetic characters. Gold threads are excluded by exact match. Within the remaining pool, candidates are drawn per source using a bucket-mode strategy: for each source the pipeline maintains four label-combination buckets (T1+T3, T1-only, T3-only, neither) and annotates items drawn from a spread-sampled pool of up to 20,000 candidates until each bucket accumulates at least 500 successful annotations or the candidate pool is exhausted. This design ensures balanced coverage of all four T1 $\times$ T3 combinations within each platform. It also explains the asymmetric thread counts relative to the assembled pool: Reddit contributes 77,455 assembled threads but far fewer silver threads.

The completed silver corpus contains **148,474 comments** in **35,431 threads**. T1 and T3 are

strongly correlated (98.9 % of T1-positive comments are also T3-positive), yet the residual mixed-label cells: norm violation without face-aggression (2.4 %) and face-aggression without norm violation (0.3 %), confirm the peripheral decoupling predicted by the multi-test framework (§2). Full label distributions and train/validation/test splits are given in Appendix A.4.

5 Method

Per-comment toxicity prediction with full thread context requires a classifier that can fit a full conversation in the context window. The longest comment thread in our conversational corpus exceeds 2,000 tokens, well beyond the 512-token ceiling of prior Hungarian encoders such as HuBERT (Nemeskey, 2021). We address this with **ModernBERT-HU-4k**, a new Hungarian encoder pretrained with a 4,096-token context window.

5.1 Tokenizer

We train a new byte-pair encoding tokenizer from scratch on the pretraining corpus. The tokenizer achieves a **fertility of 1.52** on Hungarian text, measured as subword tokens per alphabetic word on the 61 M-word silver corpus (URLs, numerals, and punctuation-only tokens excluded). On the same corpus, mBERT scores 2.23 and English ModernBERT-base scores 2.87. Lower fertility is directly relevant to toxicity: Hungarian morphological aggression vectors (T/V address shifts, sarcastic diminutive suffixes, euphemised taboo forms) are single tokens in our vocabulary, rather than being split across subwords and losing their distinctive distributional signal.

5.2 Pretraining corpus

The model is pretrained on five sources (≈ 4.4 B tokens, 1 epoch at 1,024-token context):

- **FineWeb (Hungarian subset)**: FineWeb (Penedo et al., 2024) is a high-quality filtered web text dataset providing broad lexical and topical coverage;
- **Hungarian European Parliament** (Koehn, 2005): formal register, policy and legislative domain;
- **Conversational corpus**: an earlier crawl of the same five platforms described in §4, giving the model in-domain thread structure without any labels; this crawl predates the collection of the gold comments;

- **Toxic-OSCAR**: an OSCAR-derived (Ortiz Suárez et al., 2019) subset that retains toxic-register content, ensuring the encoder develops representations that distinguish toxic from neutral discourse rather than suppressing that signal during pretraining;
- **ParlaMint-HU** (Erjavec et al., 2025): Hungarian parliamentary proceedings, supplying formal political register and politically sensitive vocabulary (a key domain for toxicity).

5.3 Training procedure

We follow the ModernBERT pretraining recipe (Warner et al., 2025), only changing the context length and extension, which was done in two steps.

Stage 1: 1k context. The full pretraining corpus is used with a 1,024-token context window. This covers the majority of individual comments and short threads and establishes strong token-level representations cheaply.

Stage 2: 4k context. The conversational corpus is repacked to 4,096-token sequences and used for continued MLM pretraining.

Gold data and pretraining. The gold comments were collected after both pretraining stages had completed, so neither stage could have seen gold text, with or without labels.

Pre-training evaluation. On the HuLU benchmark (full results in Appendix A.3), ModernBERT-HU-4k achieves overall ACC=0.691, MCC=0.398, $F_1=0.651$, leading on HuWNLI (ACC=0.617): the only task that penalises generative models for answer-format mismatch. At 149 M parameters, it is within 6–8 points of models $27\times$ (Racka-4B) and $54\times$ (PULI-LlumiX-Llama-3.1 Chat) larger, without instruction tuning.

5.4 Fine-tuning for thread-contextual toxicity classification

Task formulation. For each comment in a thread we predict two independent binary labels: T1 (face-aggression) and T3 (norm violation). The comment is encoded together with its full antecedent thread context: all preceding turns are concatenated in temporal order, separated by a [I] delimiter token. A special [TGT] token is prepended to the target comment; the sequence is then appended to the context and fed to the encoder in a single forward pass. Sequences exceed-

ing 4,096 tokens are left-truncated at turn boundaries, preserving the most recent context. A single linear layer projects the [TGT] hidden state to two logits, each independently optimized with binary cross-entropy, yielding per-label probability estimates via sigmoid activation. Full fine-tuning (FFT) hyperparameter details are given in Appendix A.10.

6 Experiments

6.1 Setup

All experiments were conducted on a single consumer-grade NVIDIA GeForce RTX 5070 Ti, demonstrating that competitive Hungarian toxicity models can be trained without access to server-grade infrastructure.

Data. Models are trained on the silver training split (31,887 threads, 135,095 comments); the silver validation split (1,772 threads) was used for hyperparameter tuning and model selection. Final evaluation uses two held-out sets: the silver test split (1,772 threads) and the gold annotated corpus (485 annotated items), which was kept blind throughout training and used exclusively for final evaluation. Because gold items carry expert labels free of LLM annotation noise, gold-set results are treated as the primary quality measure.

Metrics. For each label we report accuracy, macro F1, and AUC-ROC. Primary model selection uses F1_MIN (the lower of the two per-label macro F1 scores) to penalise configurations that improve one label at the expense of the other.

6.2 Results

Gold annotation and LLM benchmark.

Corpus-level results are reported in §4. T3 (norm violation) achieves moderate human agreement ($\alpha = 0.511$) and the strongest LLM-to-gold alignment (Gemini 3.1 Pro: $\kappa = 0.709$, $F_1 = 0.796$); T1 (face-aggression) is harder for both humans ($\alpha = 0.300$) and LLMs (best $F_1 = 0.769$ by Gemini 3 Flash).

Thread-conditioned toxicity classification. Table 4 reports ModernBERT-HU-4k full fine-tuning (FFT) results on the held-out gold test set ($n = 485$). The T1–T3 performance gap replicates the human IAA pattern (§4): norm violation is the more surface-accessible signal. For T1, class-1 precision and recall are closely matched (0.791 vs. 0.741), indicating that task-specific fine-tuning

Table 4: ModernBERT-HU-4k (FFT) on the held-out gold test set ($n = 485$; gold items with resolved labels for both T1 and T3). Cls: class label (0 = benign, 1 = toxic); P: precision, R: recall, mF1: macro F_1 ; Acc: accuracy; AUC: AUC-ROC.

	Cls	P	R	F ₁	mF1	Acc	AUC
T1	0	.841	.875	.858	.811	.823	.883
	1	.791	.741	.765			
T3	0	.897	.905	.901	.836	.862	.908
	1	.779	.764	.771			

substantially reduces the recall suppression observed in zero-shot LLMs.

Table 5 places ModernBERT-HU-4k in a unified ranking against the zero-shot LLMs. With threshold optimisation, it achieves $F_1 = 0.768$ (bootstrap 95% CI [0.723, 0.810], $n=485$, 10,000 iterations), 0.001 below Gemini 3 Flash; that model and Gemini 3.1 Pro both fall well within the interval, so point-estimate ordering among the top three carries no weight. Threshold optimisation adds $\Delta F_1 = +0.033$ over the default ($t=0.5$). ModernBERT-HU-4k is the strongest open-weight, Hungarian-native model on this benchmark, competitive with frontier proprietary systems at a fraction of the inference cost.

Inter-model agreement (Appendix A.6) shows what the model has learned: ModernBERT-HU-4k agrees most strongly with Gemini 3 Flash ($\kappa = 0.728$), ahead of every other system by a clear margin (next is Claude Sonnet 4.6 at $\kappa = 0.692$). Gemini 3 Flash is precisely the model that generated the silver labels it was trained on, so the agreement reflects a student–teacher relationship, and ModernBERT-HU-4k is best understood as a distilled approximation of Gemini 3 Flash’s annotation behaviour.

Thread context. A context-free baseline that encodes only the target comment, fine-tuned under the same recipe, loses T1 positive-class recall from 0.743 to 0.567 and T1 positive-class F_1 from 0.759 to 0.679 ($\Delta = -0.080$); T3 is almost unchanged (0.768 $\rightarrow$ 0.751, $\Delta = -0.017$; full numbers in Table 12, Appendix A.8). Because sampling is over comments, many gold targets are thread-initial (mean position 0.73), so this effect is carried by the minority of targets that have antecedent turns, which makes it more striking, not less. Face-aggression is the relational, context-dependent test; it is the one context recov-

Table 5: Unified model ranking versus human gold ($n = 485$). Avg F_1 : mean of positive-class F_1 across T1 and T3, the same quantity as the F_1 columns of Table 3 (the macro F_1 of Table 4 is a different average). **Highlighted rows**: ModernBERT-HU-4k-toxic †: Hungarian-specific model. *: supervised T2-transfer baseline (message-level, not zero-shot; scored on the 408 of 485 items it covers).

Model	Avg F_1	Avg κ
Gemini 3 Flash	.769	.655
ModernBERT-HU-4k	.768	.648
Gemini 3.1 Pro	.756	.643
Nemotron Ultra 550B	.738	.598
ModernBERT-HU-4k (FFT, $t=0.5$)	.735	.622
MiniMax M3	.705	.591
Claude Sonnet 4.6	.704	.584
GPT-OSS 120B	.680	.529
DeepSeek V4 Pro	.675	.552
Claude Opus 4.8	.663	.541
Qwen 3.7 Max	.619	.500
HuBERT-SetFit [†]	.600	.328
Racka 4B [†]	.521	.349
PULI-LlumiX-Llama-3.1 Chat [†]	.520	.400

ers. Norm violation is surface-readable and does not need the thread.

6.3 Analysis

T1/T3 difficulty asymmetry is construct-level, not artefactual. The T1/T3 performance gap is consistent across human annotators ($\alpha = 0.300$ vs. 0.511), zero-shot LLMs (best $F_1 = 0.769$ vs. 0.796), and the fine-tuned encoder ($F_1 = 0.811$ vs. 0.836, AUC = 0.883 vs. 0.908): T3 maps onto surface-accessible features (slurs, explicit threats), whereas T1 requires reasoning about relational intent within interactional context. The fine-tuned model’s positive-class F_1 exceeds the inter-annotator baseline on both tasks (T1: 0.765 > 0.564; T3: 0.771 > 0.664), reflecting multi-annotator resolved labels rather than superhuman performance.

Where the T1 recall gain comes from. Zero-shot LLMs are conservative on T1: precision is high but recall is suppressed, since models default to benign on borderline implicational impoliteness (Cercas Curry et al., 2024). The silver labels inherit this bias, so the fine-tuned model gains no recall from more complete supervision. The gain (T1 class-1 recall 0.741, against 0.688 for Gemini 3 Flash, the model that produced the silver labels) comes instead from a decision-boundary correction on top of the distilled representation: a positive-class weight $w_{T1}^+ = 2.95$, close to

the inverse of the 25.3% T1 base rate, plus per-label threshold tuning on the silver validation split ($\Delta F_1 = +0.033$ over $t = 0.5$).

Task-specific fine-tuning compensates for scale.

ModernBERT-HU-4k (149 M parameters) outperforms ten of eleven zero-shot LLMs and is within $\Delta F_1 = 0.001$ of the eleventh, despite being fine-tuned on a single consumer GPU (RTX 5070 Ti, 16 GB). The prior Hungarian T2 classifier (HuBERT-SetFit) ranks below every zero-shot LLM except the two Hungarian-specific ones, confirming that the encoder’s gains are driven by T1/T3-specific supervision rather than Hungarian-language pretraining alone.

7 Conclusion

We introduced a pragmatically grounded three-test framework for Hungarian online toxicity, constructed a large-scale silver-annotated conversational corpus with a gold-annotated pilot subset, and developed ModernBERT-HU-4k, a new long-context Hungarian encoder fine-tuned for thread-contextual toxicity classification. The framework addresses the conceptual deficit identified in the only prior Hungarian work (Hatvani and Yang, 2025) and instantiates the relational research agenda of Amironesei and Díaz (2023) for a non-English, morphologically rich language.

Limitations

The gold-annotated corpus is small (485 items) and drawn from a limited set of Hungarian platforms; the silver corpus is larger but inherits noise from LLM annotation. Results may not generalise to all registers of Hungarian online discourse. The three-test framework requires annotators with deep pragmatic understanding, which raises costs compared to binary labelling. Cross-cultural transfer experiments are limited by the absence of comparable non-English CRAFT-style corpora. The thread-level context-trigger field (M -indices of foreshadowing turns, §4) is collected in the gold corpus but is not consumed by any model or analysis in this submission; at 6 % coverage (32 of 498 items) the field is too sparse for reliable modelling and is reserved for future causal analysis of conversational derailment trajectories.

Ethics Statement

Data collection. All data were collected from publicly accessible comment sections under each

platform’s terms of service for research use. Personally identifiable information (names, usernames, e-mail addresses, URLs) was removed prior to annotation following the pseudoanonymisation procedure of [Hatvani et al. \(2023\)](#); public figures were excluded from pseudoanonymisation as their public-sphere activity carries no heightened privacy risk. Illustrative examples in this paper are synthetic reconstructions of corpus patterns, not verbatim posts; the released corpus itself retains verbatim (pseudonymised) text, since the morphological markers under study do not survive paraphrase.

Human annotation. Annotators were adult Hungarian-speaking volunteers (ages 28–42, gender-balanced). Participation was voluntary; participants could withdraw at any time without consequence. Given exposure to toxic and hateful content, psychological consultation was offered to all annotators throughout the project. Annotation was fully anonymous: personal and demographic data were collected on a separate, unlinked form and are not included in the released corpus. The dataset records only irreversible cryptographic hashes as annotator identifiers, which cannot be mapped back to real user identities.

Dual-use and release. The released models and corpus carry a dual-use risk. Distribution will be under a research-only licence with usage guidelines against adversarial misuse (e.g. targeted harassment, filter evasion).

Acknowledgments

References

- Razvan Amironesei and Mark Díaz. 2023. [Relationality and offensive speech: A research agenda](#). In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*, pages 85–95. Association for Computational Linguistics.
- Roger Brown and Albert Gilman. 1960. The pronouns of power and solidarity. In Thomas A. Sebeok, editor, *Style in Language*, pages 253–276. MIT Press, Cambridge.
- Michael Castelle. 2018. [The linguistic ideologies of deep abusive language classification](#). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.
- Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. [Subjective isms? on the danger of conflating hate and offence in abusive language detection](#). In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 275–282, Mexico City, Mexico. Association for Computational Linguistics.
- Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. [Trouble on the horizon: Forecasting the derailment of online conversations as they develop](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.
- Zsolt Csibi, Bence György Gortka, Natabara Gyöngyössi, Kornél Nagy, Dávid Márk Nemeskey, Martin Sallai, András Simonyi, András Márk Szekeres, and Gábor Palkó. 2026. [Racka: Efficient hungarian llm adaptation on academic infrastructure](#). Preprint, arXiv:2601.01244.
- Jonathan Culpeper. 2011. *Impoliteness: Using Language to Cause Offence*. Number 28 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge.
- Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Korzhinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Dargis, and 18 others. 2025. [ParlaMint II: Advancing comparable parliamentary corpora across Europe](#). *Language Resources and Evaluation*, 59(3):2071–2102.
- European Parliament and Council of the European Union. 2022. [Regulation \(EU\) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services \(Digital Services Act\)](#). Official Journal of the European Union, L 277/1.
- Péter Hatvani, László János Laki, and Zijian Győző Yang. 2023. [A pseudonymization tool for Hungarian](#). *Annales Mathematicae et Informaticae*, 58:69–80.
- Péter Hatvani and Zijian Győző Yang. 2025. [Automated detection of toxic comments in Hungarian](#). *Annales Mathematicae et Informaticae*, 61:108–117.
- Michael Hoffmann, Jan Fillies, Jophin John, Antonis Maronikolakis, Ajay Navilarekal, Sahana Udupa, and Nicolay Hammer. 2025. [Temporal robustness in hate speech detection: Updating German classifiers with advanced AI infrastructures](#). In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pages 1171–1178.

- Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. [Hate speech criteria: A modular approach to task-specific hate speech definitions](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Philipp Koehn. 2005. [Europarl: A parallel corpus for statistical machine translation](#). In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.
- Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. [In data we trust: A critical analysis of hate speech detection datasets](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.
- Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. [A modular taxonomy for hate speech definitions and its impact on zero-shot LLM classification performance](#). In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 490–521, Vienna, Austria. Association for Computational Linguistics.
- Microsoft Corporation. 2024. [Playwright](#). Version 1.x.
- Peter Mistahl. 2026. [Lingua: The most accurate natural language detection library for python](#).
- Dávid Márk Nemeskey. 2021. [Introducing huBERT](#). In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021)*, pages 3–14.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. [Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures](#). In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMC-7) 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. [Toxicity detection: Does context really matter?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). *Preprint*, arXiv:2406.17557.
- Raquel Recuero. 2024. [The platformization of violence: Toward a concept of discursive toxicity on social media](#). *Social Media + Society*, 10(1).
- Gal Ron, Effi Levi, Odelia Oshri, and Shaul R. Shenhav. 2023. [Factoring hate speech: A new annotation framework to study hate speech in social media](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 215–220, Toronto, Canada. Association for Computational Linguistics.
- The PRAW Development Team. 2024. [PRAW: Python Reddit API Wrapper](#). Version 8.0.2.
- Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. [From languages to geographies: Towards evaluating cultural bias in hate speech datasets](#). In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.
- Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. [Efficient few-shot learning without prompts](#). *arXiv preprint arXiv:2209.11055*.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. [Ex machina: Personal attacks seen at scale](#). In *Proceedings of the 26th International Conference on World Wide Web, WWW '17*, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Zijian Győző Yang, Ágnes Bánfi, Réka Dodé, Gergő Ferenczi, Flóra Földesi, Péter Hatvani, Enikő Héja, Mariann Lengyel, Gábor Madarász, Mátyás Osváth, Bence Sárossy, Kristóf Varga, Tamás Váradi, Gábor Prószéky, and Noémi Ligeti-Nagy. 2025. [Chat-PULI: Enhancement to the first Hungarian conversational model](#). *Annales Mathematicae et Informaticae*, 61:261–274.
- Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. [Conversations gone awry: Detecting early signs of conversational failure](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

A.1 Corpus Construction and Annotation Pipeline

Figure 1 summarises the end-to-end pipeline, from raw multi-platform comment collection through cleaning, anonymisation, language gating, and weak toxicity labelling; gold annotation with inter-annotator agreement checks and third-round escalation; zero-shot LLM benchmarking and teacher selection; and silver-label propagation used to fine-tune and evaluate the final encoder.

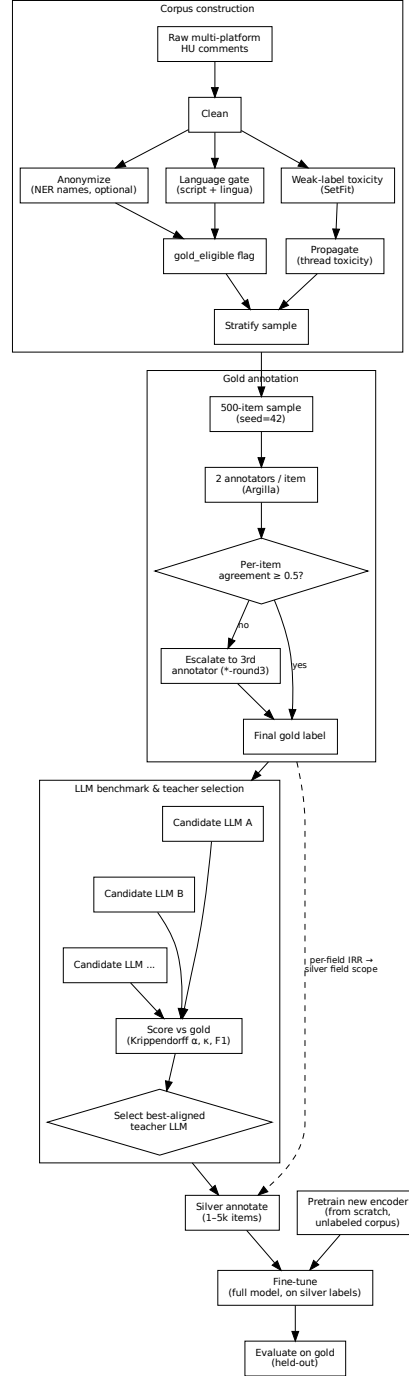


Figure 1: Corpus construction, gold annotation, LLM teacher selection, and silver-labelled fine-tuning pipeline.

Table 6: Source-level corpus statistics after pseudoanonymisation, thread assembly, and weak labelling with the HuBERT-SetFit toxic-comment classifier (Hatvani and Yang, 2025). N_{comm} : total comments; n_{tox} : comments flagged toxic; N_{thr} : threads after assembly; $n_{\text{thr-tox}}$: threads containing at least one toxic comment. Sources are ordered by decreasing comment-level toxicity rate.

Source	N_{comm}	n_{tox}	% toxic	N_{thr}	$n_{\text{thr-tox}}$	% tox thr
Mandiner	12,909	9,942	77.0	2,510	2,445	97.4
Kuruc	119,207	88,084	73.9	1,885	1,873	99.4
Reddit	2,264,061	1,158,293	51.2	77,455	61,680	79.6
YouTube	868,023	407,872	47.0	6,793	6,624	97.5
TikTok	297,849	72,501	24.3	9,525	6,213	65.2
Total	3,562,049	1,736,692	48.8	98,168	78,835	80.3

Table 7: Full HuLU benchmark results per task. A = ACC, M = MCC, F = F_1 . Best per task across all models in **bold**.

Task	ModernBERT-HU-4k			PULI-LlumiX-Llama-3.1 Chat			Qwen3-4B		
	A	M	F	A	M	F	A	M	F
HuCOLA	.885	.648	.823	.899	.692	.897	.811	.348	.784
HuCOPA	.690	.380	.690	.936	.872	.936	.559	.118	.558
HuSST	.701	.416	.606	.780	.560	.770	.752	.502	.743
HuRTE	.634	.258	.629	.898	.794	.898	.908	.814	.908
HuWNLI	.617	.237	.536	.380	-.282	.367	.503	-.098	.386
HuCommitmentBank	.621	.453	.624	.485	.274	.459	.738	.608	.732
Overall	.691	.398	.651	.729	.485	.721	.711	.382	.685

Task	Qwen3-4B-Base			Racka-4B		
	A	M	F	A	M	F
HuCOLA	.825	.404	.803	.862	.566	.856
HuCOPA	.585	.171	.584	.799	.600	.799
HuSST	.754	.508	.751	.760	.514	.751
HuRTE	.887	.772	.887	.879	.755	.879
HuWNLI	.537	-.060	.407	.567	.103	.455
HuCommitmentBank	.629	.473	.611	.639	.474	.637
Overall	.702	.378	.673	.751	.502	.730

A.2 Source-Level Corpus Statistics

Table 6 reports the full per-source breakdown of comment and thread counts after pre-screening.

A.3 HuLU Per-Task Breakdown

Table 7 shows per-task results for all five models on the HuLU benchmark. ModernBERT-HU-4k leads only on HuWNLI; on all other tasks the generative models are stronger, consistent with the expectation that encoder-only models are disadvantaged on tasks that reward free-form completion.

A.4 Silver Corpus Label Distribution

Table 8 reports the comment-level *combo_cat* counts across data splits. Stratification by *combo_cat* keeps all four label combinations within 1.5 percentage points of their global proportion across all three splits, confirming balanced coverage. Per-label marginals: T1 positive 37,632 (25.3 %); T3 positive 40,808 (27.5 %).

Table 9: Model identifiers and access details for the zero-shot LLM benchmark. †: Hungarian-specific model run locally.

Model (paper name)	Model identifier	Provider
Claude Opus 4.8	anthropic/claude-opus-4.8	amazon-bedrock/us
Claude Sonnet 4.6	anthropic/claude-sonnet-4.6	google-vertex/europe
DeepSeek V4 Pro	deepseek/deepseek-v4-pro	baidu/fp8
Gemini 3 Flash	google/gemini-3-flash-preview	google-vertex/global
Gemini 3.1 Pro	google/gemini-3.1-pro-preview	google-ai-studio
GPT-OSS 120B	openai/gpt-oss-120b	dekallm/bf16
MiniMax M3	minimax/minimax-m3	novita
Nemotron Ultra 550B	nvidia/nemotron-3-ultra-550b-a55b	nvidia
Qwen 3.7 Max	qwen/qwen3.7-max	alibaba
PULI-LlumiX-Llama-3.1 Chat [†]	NYTK/PULI-LlumiX-Llama-3.1	HuggingFace local
Racka 4B [†]	elte-nlp/Racka-4B	HuggingFace local

Table 8: Comment-level *combo_cat* distribution by split (counts). Stratification is by thread; comments within a thread are never split across partitions.

Split	no_no	yes_yes	no_yes	yes_no	Total
Train	97,488	33,972	3,262	373	135,095
Val	4,793	1,569	166	17	6,545
Test	4,973	1,679	160	22	6,834
Total	107,254	37,220	3,588	412	148,474
(%)	72.2	25.1	2.4	0.3	100

A.5 Zero-Shot LLM Benchmark: Model Identifiers

Table 9 lists the exact model identifiers for all eleven zero-shot benchmark runs. API models were accessed via OpenRouter (openrouter.ai/api/v1) with temperature=0 and allow_fallbacks=false pinned to the provider listed in column 3. The two Hungarian-specific models were run locally via HUGGINGFACE Transformers with greedy decoding (temperature=0).

A.6 Inter-Model Agreement

Table 10 reports pairwise Cohen’s κ between ModernBERT-HU-4k (FFT) and each zero-shot LLM, averaged across T1 and T3.

Table 10: ModernBERT-HU-4k (FFT) inter-model agreement with zero-shot LLMs (average Cohen’s κ across T1 and T3). †: Hungarian-specific model.

Model	κ
Gemini 3 Flash	.728
Claude Sonnet 4.6	.692
Gemini 3.1 Pro	.681
Nemotron Ultra 550B	.673
MiniMax M3	.638
DeepSeek V4 Pro	.588
Claude Opus 4.8	.586
GPT-OSS 120B	.551
Qwen 3.7 Max	.530
PULI-LlumiX-Llama-3.1 Chat [†]	.378
Racka 4B [†]	.316

A.7 Disagreement Analysis for the Gold Corpus

Table 11 breaks down inter-annotator disagreement on the gold corpus by platform, for aggression detection (T1) and mechanism classification (Dim3).

Table 11: Inter-annotator disagreement on aggression detection (T1) and mechanism classification (Dim3), by platform

Platform	Aggression (T1)			Mechanism (Dim3) ^a		
	<i>N</i>	Disagree	Rate (%)	<i>N</i>	Disagree	Rate (%)
kuruc	97	54	55.7	38	21	55.3
mandiner	98	44	44.9	39	21	53.8
reddit	94	29	30.9	14	8	57.1
tiktok	96	17	17.7	5 ^b	4	80.0
youtube	100	26	26.0	15	11	73.3
Total	485	170	35.1	111	65	58.6

^a Mechanism is a conditional field, answered only

when an annotator marked `t1_aggression = yes`; *N* is restricted to items where ≥ 2 annotators both answered it, hence the smaller counts.

^b *N* = 5; interpret this cell with caution.

A.8 Thread-Context Ablation

Table 12 reports the full thread-context ablation on the held-out gold set, comparing the model with and without preceding-turn context.

Table 12: Thread-context ablation on the held-out gold set ($n = 485$). cls-1 F_1 : positive-class F_1 ; mF1: macro F_1 ; AUC: AUC-ROC.

Setting	T1 aggression			T3 norm violation		
	cls-1 F_1	mF1	AUC	cls-1 F_1	mF1	AUC
Full context	.759	.807	.881	.768	.834	.908
No context	.679	.767	.853	.751	.829	.907
Δ	-.080	-.040	-.028	-.017	-.005	-.001

A.9 LLM Annotation Prompt (v2)

Warning: the examples below include offensive language. The annotation prompt covers T1 and T3 only; secondary dimensions (Dim II–V) are omitted because their LLM–human IAA fell below the reliable-labelling threshold (§4). The prompt is written in Hungarian to match the target language. All models were called at temperature = 0 for reproducibility.

Hungarian original: The prompt numbers jump from 1) to 3) because T2 (antisocial perlocutionary effect) is not annotated by human raters, it is derived from thread structure by the pipeline, and is therefore omitted from the annotation instructions.

Magyar online kommentek pragmatikai-relációs toxicitását annotáld, egy nyelvészeti kutatási gold korpusz-kompatibilis ezüstkorpuszához. NEM egyes szavakat címkézel "csúnya / nem csúnya" alapon: azt ítéled meg, hogy egy üzenet a SAJÁT beszélgetési környezetében toxikus-e.

ALAPELVEK

- Mindig CSAK a [>>] jellel megjelölt utolsó hozzászólást (M_n) értékeled.
- A korábbi fordulók (M₀, M₁, ...) kizárólag kontextus -- ket nem annotáld.
- Az üzenetet SOHA ne ítéld meg önmagában, mindig a megelőző fordulók fényében.
- A trágárság NEM azonos a toxicitással. Sok toxikus üzenetben egy csúnya szó sincs (gúny, "toxikus pozitívítás", kirekesztés); sok trágár üzenet baráti ugratás, tehát nem toxikus.

A TOXICITÁS KÉT ANNOTÁLANDÓ TESZTJE

Mindkettő külön, igen/nem + magabiztosság (1-4) skálán ítéld meg.

1) T1 -- AGRESSZIÓ

Tartalmaz-e az üzenet konvencionált bántó formulát (sértés, fenyegetés, lekezelés, elhallgattatás, átok) vagy nyílt személyes támadást a CÍMZETT ARCA -- önképe, jelleme, kompetenciája, csoporthovatartozása -- ellen, nem a véleménye, hanem a SZEMÉLYE ellen?

- Igen: "Te agyhalott barom, fogd be." / "Kapt be, idióta."
- Nem (csak véleménynyilvánítás): "Ez az érved hibás, mert..." (vita, nem agresszió)

3) T3 -- NORMASZEGÉS

Megsérti-e az üzenet az adott KÖZEG (hírközlő, fórum, videóközlő, csoport) elvárta, ill. nyelvi viselkedését EBBEN a pillanatban, egy magyar anyanyelvű, a közeg ismerője szerint?

Ugyanaz a mondat egy gamer-fórumon norma, egy gyászoló posztban súlyos normaszegés.

FONTOS: T2 (antiszocialitás) NEM a te dolgod -- azt a rendszer számítja a szálszerkezetben. Ne állítsd el, ne tippeld meg. Semmilyen más mezt NE adj meg a JSON-ban.

KÖZPONT ÉS PERIFÉRIA

- Mag eset: mindkét teszt "igen" (pl. nyílt sértés, csoportos uszítás).
- Csak T3 (T1 nem): "toxikus pozitivitás" -- meleg, biztató megfogalmazás, amely valós nehézséget hallgattat el, hibáztat vagy leértékel -> TOXIKUS.
- Csak T1 (T3 nem): kölcsönös, a közeg normájának megfelelő baráti ugratás (banter), elhallgattató hatás nélkül -> NEM toxikus.

NEHÉZ ESETEK

- Irónia/szarkazmus: a felszíni pozitív szó nem dönt. Tartható-e a jóhiszem olvasat a szálban? Ha nem -> T1/T3 igen.
- Baráti trágárság (banter): kölcsönös + közeg normája + nincs elhallgattatás -> T1 lehet igen, de T3 nem -> NEM toxikus.
- Toxikus pozitivitás: meleg felszín + nehézség elhallgattatása -> T3 igen, T1 nem.
- Eufemizált tabu (ba*zeg 'f**k', wa*ze 'f**k'): ugyanúgy számít, mint a teljes alak.
- Politikai üres jelöl (Soros, migráncs, libsi): önmagában NEM mindig toxikus; a kontextus és a szándék dönt.

MAGABIZTOSSÁG (1-4): 1 = nagyon bizonytalan, 4 = teljesen biztos.

Kétértelm. esetben adj 1-2-t.

KIMENET -- válaszolj KIZÁRÓLAG egyetlen JSON objektummal:

```
{
  "t1_aggression": true | false,
  "t1_confidence": 1 | 2 | 3 | 4,
  "t3_norm_violation": true | false,
  "t3_confidence": 1 | 2 | 3 | 4
}
```

Csak a JSON objektumot add vissza. Ne tegyél köré ```-jelet vagy magyarázatot.

English translation (paraphrase): Task. Annotate the pragmatic-relational toxicity of Hungarian online comments for a gold-corpus-compatible silver dataset. Do not label individual words as "offensive / not offensive": judge whether a message is toxic *in its own conversational context*.

Core principles. (1) Evaluate only the last turn (marked M_n with [»]); earlier turns are context only, not annotated. (2) Never judge the message in isolation; always consider it in the light of preceding turns. (3) Profanity $\neq$ toxicity: many toxic messages contain no offensive words (sarcasm, exclusion, "toxic positivity"); many profane messages are friendly banter and therefore not toxic.

T1: Aggression. Does the message contain a conventionalised offensive formula (insult, threat, condescension, silencer, curse) or explicit personal attack on the *face* of the addressee (their self-image, character, competence, or group membership), targeting them as a *person*, not their opinion?

T3: Norm violation. Does the message violate the expected, appropriate linguistic behaviour of the *platform context* (news comment, forum, video comment, group) at this moment, according to a native Hungarian speaker familiar with the community? The same sentence may be the norm on a gaming forum and a serious violation on a bereavement post.

Important. T2 (antisociality) is *not* the model's task; it is derived from thread structure by the pipeline.

Centre and periphery. (a) Core case: both tests fire (overt insult, group incitement). (b) T3 only, not T1: "toxic positivity": a warm, encouraging message that silences or blames a real difficulty $\Rightarrow$ TOXIC (T3 yes, T1 no). (c) T1 only, not T3: mutual banter conforming to community norms with no silencing

effect $\Rightarrow$ NOT toxic (T3 no).

Output. Return exactly one JSON object with keys `t1_aggression`, `t1_confidence`, `t3_norm_violation`, `t3_confidence` and no other fields.

A.10 Pretraining and Fine-Tuning Hyperparameters

Stage 2 pretraining hyperparameters are given in §5; Stage 1 follows the ModernBERT recipe (Warner et al., 2025) (MLM, 30% mask ratio, ≈ 4.4 B tokens, 219,621 steps). Full fine-tuning (FFT) updates all model parameters. Model selection is performed on the silver validation split using F1_MIN (the lower of the two per-label macro F1 scores).

Table 13: ModernBERT-HU-4k full fine-tuning (FFT) hyperparameters.

Hyperparameter	FFT
Learning rate	3.23×10^{-5}
Batch size (per device)	16
Gradient accumulation	2
Warmup ratio	0.030
Epochs	5
Optimizer	AdamW (8-bit)
Adam ε	10^{-6}
Weight decay	0.084
Head dropout	0.012
w_{T1}^+	2.95
w_{T3}^+	2.64
Selection metric	F1_MIN