From News Streams to Narrative Intelligence Briefs: LLM-Assisted Political Discourse Analysis in the Hungarian 2026 Pre-Election Context

Ekaterina Loginova, Maksim Ermakov, Stephan Khramov
Aletheos
kate@aletheos.team, maksim@aletheos.team, stephan@aletheos.team

Abstract

Journalists, activists and policymakers need scalable ways to convert high-volume political news into actionable narrative intelligence, yet most NLP pipelines stop at classification outputs that are difficult to operationalise. We present a case study assessing whether LLMs, constrained by explicit analytical schemas and multi-stage validation, can reliably produce structured narrative intelligence briefs from Hungarian pre-election news. Our pipeline processes 574 election-relevant articles from 21 sources through three stages: (1) per-article extraction of Narrative Policy Framework event frames and SemEval-taxonomy manipulation techniques; (2) embedding-based narrative clustering; and (3) constrained brief generation with five sections—narrative summary, character map, manipulation profile, escalation assessment, and counter-strategy—where counter-strategies are grounded in curated external sources. Evaluation through human experts and LLM-as-judge reveals a consistent quality gradient: descriptive sections outperform prescriptive ones, with failures driven by contextual insufficiency rather than hallucination. We document failure modes and assess which components support semi-automation.

Keywords: narrative intelligence, political discourse analysis, large language models, propaganda detection, Narrative Policy Framework, election monitoring, human-in-the-loop NLP

1. Introduction

News monitoring teams face a volume–capacity mismatch. For under-resourced languages, qualified analysts are even scarcer—the Atlantic Council's DFRLab reports that analysing pro-Kremlin narratives required four Russian-fluent researchers working for months (Brookie and Digital Forensic Research Lab, 2023). Our working corpus of 574 election-relevant Hungarian articles would require ∼15 hours of reading alone, whereas reaction is often expected within minutes.

Practitioners need more than topic detection or claim classification; they need to understand the *narrative landscape*: who is convinced of what, by whom, through which mechanisms—and what responsible counter-strategies exist (UK Government Communications Service, 2019). Two NLP subfields address parts of this: propaganda technique detection (Da San Martino et al., 2019; Piskorski et al., 2023) identifies *how* content manipulates; narrative extraction via the Narrative Policy Framework (NPF; Jones and McBeth, 2010; Shanahan et al., 2018) identifies *who* is cast in which role. Yet neither produces *structured intelligence briefs* integrating analysis with actionable counter-strategies. Meanwhile, practitioner frameworks for counter-propaganda—GAMMMA+ (Silverman et al., 2016), inoculation theory (van der Linden et al., 2017; Roozenbeek et al., 2022), and the counter- vs. alternative narrative distinction (Radicalisation Awareness Network, 2022)—remain manual and unintegrated with computational pipelines.

We bridge these worlds, asking: how reliable are LLM-generated narrative intelligence briefs when extraction is grounded in NPF and SemEval taxonomies, and counter-strategies are constrained by evidence-based frameworks and verified external sources? We contribute: (1) a reproducible pipeline from raw news to structured briefs[1]; (2) a counter-strategy schema operationalising practitioner frameworks within constrained generation; (3) empirical evaluation with documented failure modes; and (4) a critical assessment of evaluation methodology, including LLM-as-judge limitations.

Our setting—the Hungarian 2026 pre-election period—provides a stress test: a polarised media ecosystem with significant state-media influence. We scope this as a methodology demonstration evaluating analyst-support capabilities, not automated decision-making.

2. Related Work

The NPF provides formal structure for political narratives: characters assigned roles (hero, villain, victim), causal relationships, and implied solutions (Jones and McBeth, 2010; Shanahan et al., 2018). Gehring and Grigoletto (2023) formalised a character-role framework for computational analysis, while Otmakhova and Frermann (2025) define

[1] Repository: https://github.com/aletheos-ngo/politicalNLP2026

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 1–16
16 May 2026, Palma, Mallorca (Spain).

computational narrative structures in political news. We adopt NPF because its extractable components map directly to intelligence brief fields.

The SemEval shared tasks (Da San Martino et al., 2019; Piskorski et al., 2023; **?**) established fine-grained manipulation technique taxonomies with multilingual benchmarks. While per-sentence classification is well-studied, aggregation into narrative-level *manipulation profiles* remains underexplored. Our work addresses this gap.

LLMs show promise for political text analysis (Ziems et al., 2024; Törnberg, 2023) but exhibit partisan bias (Feng et al., 2023), hallucination (McKenna et al., 2023), and cultural flattening on non-English discourse. Counter-propaganda research warns that messaging must engage on a narrative level beyond facts alone (Carnegie Endowment for International Peace, 2024)—meaning LLMs generating strategic advice must do more than retrieve facts, yet must not fabricate. Our pipeline addresses this by separating unconstrained generation (narrative analysis grounded in source text) from constrained generation (counter-strategies bounded by external evidence).

Counter-propaganda frameworks. The UK RESIST toolkit provides a practitioner-oriented schema for disinformation reporting: narrative identification, manipulation analysis, audience assessment, and recommendations (UK Government Communications Service, 2019). The EU's RAN distinguishes counter-narratives (directly rebutting extremist content) from alternative narratives (positive reframing that undermines assumptions without engaging them), noting that alternatives must introduce "something novel" rather than generic values (Radicalisation Awareness Network, 2022). The GAMMMA+ framework structures campaign design around Goal, Audience, Messenger, Message, Media, and Action (Silverman et al., 2016). Inoculation research— preemptive exposure to weakened manipulation—has shown scale effectiveness, including prebunking campaigns in Central and Eastern Europe (Roozenbeek et al., 2022). Moral reframing demonstrates that arguments couched in the target audience's values are more persuasive than those using the communicator's values (Feinberg and Willer, 2015). Our counter-strategy schema operationalises these frameworks: each brief's response section is structured around the counter- vs. alternative narrative distinction, grounded in inoculation principles, and constrained to propose specific messengers and rebuttals rather than generic advice.

3. Pipeline Overview

The pipeline has three stages, each independently auditable. **Stage 1** (Extraction) processes each article twice: once for NPF frame extraction (actor–role–action–target with causal claims) and once for SemEval-taxonomy manipulation technique detection. **Stage 2** (Clustering) embeds extracted frames, clusters them into narrative groups, and assigns thematic domains. **Stage 3** (Generation) produces a five-section intelligence brief per cluster, with counter-strategies constrained by curated contextual cards. Formally: extraction $f_{GPT-4o}(articles) \rightarrow$ frames + techniques; clustering $g(frames) \rightarrow$ clusters; generation $h_{Claude}(cluster, card, principles) \rightarrow$ brief. The brief-generating model receives only extraction outputs and analyst-curated context—never raw articles, preventing selective out-of-context quoting.

4. Stage 1: Extraction

4.1. Data Collection

We collect Hungarian-language political news via RSS feeds from 21 sources spanning the media spectrum: pro-government (Magyar Nemzet, Magyar Hírláp, Mandiner, Pesti Srácok, Híradó, Demokrata), independent (Telex, Index, 444.hu, HVG, Átlátszó, Direkt36), opposition-aligned (Népszava, Magyar Narancs, Mérce, 168.hu, Hang.hu), and general portals (24.hu, ATV, Portfolio, Hírstart). Full-text extraction uses `newspaper4k`. Successive collection runs (February 2026) yield 3,708 unique articles after URL-based deduplication. A GPT-4o[2] binary classifier identifies election-relevant content, yielding **574 articles**. This is single-pass classification with no inter-annotator validation.

4.2. NPF Frame Extraction

Each article undergoes structured extraction producing elements with seven core fields: *actor* (surface form), *portrayed_role* (hero, villain, victim, narrator, neutral, unclear), *narrative_function* (claim, evaluation, threat, mobilisation, warning, policy_promise), *action_or_claim*, *target* (surface form or null), *portrayed_target_role* (adding `threat` and `proxy` to capture entities framed as abstract dangers or stand-ins—important in Hungarian rhetoric where "Brussels," "war," and "migration" frequently serve these functions), and a nested *causal_claim* (text, causal marker, relation type). Extraction uses GPT-4o (`temperature=0`, Pydantic structured output), producing **4,266 elements** across 574 articles (mean: 7.4/article).

[2]All LLM calls use GPT-4o unless stated otherwise.

The extraction prompt underwent four iterations reflecting a key design insight: **extraction reliability outperforms maximum coverage**. Version 1 (12-field schema) produced well-grounded but analytically flat extractions with ∼40% background pollution. Version 2 expanded to 15 fields but the model ignored new fields due to prompt signal-to-noise: 11 instruction blocks at the same level caused the model to follow easy constraints while silently dropping hard ones. Version 4 resolved this through: (1) schema-first layout with inline comments; (2) a single worked example with reasoning—the highest-leverage change. Full prompts are provided in Appendix D.

4.3. Manipulation Technique Detection

A parallel pass identifies rhetorical manipulation using the SemEval 2023/SlavicNLP 2025 taxonomy: 6 categories, 25 fine-grained labels (Piskorski et al., 2023). Three prompt versions progressed from F1=0.008 (zero-shot, custom taxonomy) through F1=0.104 (SemEval labels, 4 examples) to **F1=0.227** (12 language-specific examples + disambiguation table + multi-label instruction + hard vocabulary constraint), comparable to the best shared-task system (macro F1=0.21; Piskorski et al. 2025).

Error analysis revealed five systematic causes: (a) *phantom labels*—the model generated plausible but out-of-taxonomy labels (`scapegoating`: 14 FP), accounting for 24 false positives; (b) *span size mismatch*; (c) zero recall on 12/25 categories; (d) systematic label confusions; and (e) *multi-label blindness*—51% of gold annotations were multi-label but the model rarely produced overlapping labels. Per-technique performance varies substantially: techniques with clear surface cues (`guilt_by_association`, F1=0.500) outperform those requiring pragmatic inference (`obfuscation`, F1=0.000).

For brief generation, the 25 techniques are mapped to 6 intent-based categories designed for counter-strategy selection: *legitimisation*, *emotional manipulation*, *identity polarisation*, *simplification*, *distraction*, and *mobilisation*. Each technique is further annotated with emotion mappings (emotions leveraged, provoked, and cognitive effect), enabling the brief to characterise the emotional architecture of each narrative cluster (Appendix C).

4.4. Prompt Engineering Lessons

Both extraction tasks yielded generalisable findings. **Schema reduction improves quality**: dropping 8 fields improved the remaining fields' quality. **Decision tables outperform prose**: compact lookup structures reduced attention cost. **Worked examples are highest-leverage**: for NPF extraction, one example resolved schema compliance; for manipulation detection, expanding from 4 to 12 examples drove the largest F1 improvement. **Hard vocabulary constraints are essential**: the model generated plausible out-of-taxonomy labels despite explicit instructions; mitigation required both in-prompt constraints and post-processing normalisation.

5. Stage 2: Narrative Clustering

Each extracted frame is serialised preserving NPF structure—e.g., "Orbán Viktor (as hero) claims Brussels pressures Hungary about Brüsszel (as villain) because EU sanctions threaten sovereignty"—and embedded using `paraphrase-multilingual-mpnet-base-v2`. UMAP reduces dimensionality (`n_neighbors=15`, `n_components=5`, cosine metric) and HDBSCAN clusters (`min_cluster_size=15`, `min_samples=5`), yielding **80 clusters** (silhouette: 0.600, 26.4% noise).

We compared three strategies: *holistic* (full frame text), *component-based* (four NPF dimensions clustered independently), and *hybrid*. Cross-strategy agreement was near-perfect (ARI=1.000), indicating holistic clustering dominates. The component-based strategy provides analytical enrichment: it enables cross-domain archetype analysis, e.g., detecting that the "defender vs. external threat" archetype appears across EU-relations, security, and migration domains with different actors but identical role structures. We adopt the component-based strategy for its richer analytical output.

Each cluster is auto-labelled from modal actor, target, and role values, with *contested characterisation* detection when an actor appears in multiple roles across sources. Clusters are mapped to thematic domains via keyword matching against a curated bilingual dictionary (7 domains, actor matches weighted 3×). As a comparative baseline, BERTopic applied to the same corpus confirmed a fundamental distinction: BERTopic recovers *topics* (what articles are about) but not *narratives* (how stories are framed, with actor–role assignments and causal logic); see Appendix B.

Table 1 shows the five most frequent narrative structures.

6. Stage 3: Brief Generation

For each cluster, a Narrative Intelligence Brief is generated with five sections (Table 2), using Claude Sonnet[3] (`temperature=0.3`). The

[3]Brief generation uses Claude Sonnet (`claude-sonnet-4-5-20250929`) for its longer context window and structured JSON output.

	Actor→Target	A.	T.	n
1	Orbán V.→Brüsszel	hero	vill.	69
2	Trump→Orbán V.	hero	hero	45
3	Orbán V.→Tisza P.	hero	proxy	18
4	Orbán V.→Ukrajna	hero	vill.	13
5	Orbán V.→háború	hero	threat	10

Table 1: Top-5 narrative clusters by frequency. A.=actor role; T.=target role; vill.=villain. Elements with null targets excluded.

system prompt defines an analyst role, six de-escalation principles drawn from inoculation research (van der Linden et al., 2017) and WHO risk communication guidance (World Health Organization, 2017), and evaluative language boundaries per section (strictly neutral for Section A, analytical for B–C, risk-assessment for D, prescriptive-but-constrained for E).

The pipeline's central design principle separates unconstrained generation (Sections A–D: narrative analysis grounded in source text, where the model may synthesise freely) from constrained generation (Section E: counter-strategies bounded by contextual cards and de-escalation principles, where the model must not draw on parametric knowledge for facts).

Before generation, articles are filtered by semantic coherence with the cluster centroid (cosine threshold 0.35), and data is aggregated: up to 15 diverse frames, manipulation technique frequencies with emotion profiles, representative excerpts, and source classification metadata.

Section	Content and constraints
A. Summary	Actors, claims, causal logic. Strictly neutral.
B. Characters	Hero/villain/victim/contested with Hungarian-language evidence and source attribution. No inferred roles.
C. Manipulation	Top techniques across cluster; master narrative classification; emotional levers. Must use taxonomy labels with evidence.
D. Escalation	Monitor/Concern/High-Risk with specific signals and trajectory assessment.
E. Counter-strategy	Structured responses (see below).

Table 2: Narrative Intelligence Brief schema.

6.1. Counter-Strategy Generation: Frameworks and Constraints

The counter-strategy section is the most novel and most risky component of the brief. To prevent the LLM from producing generic or unsupported advice, we constrain generation along three axes: (a) a structured response typology drawn from counter-propaganda research; (b) contextual cards providing verified country-specific information; and (c) de-escalation principles embedded in the system prompt. Drawing on the RAN counter- vs. alternative narrative distinction (Radicalisation Awareness Network, 2022) and GAMMMA+ (Silverman et al., 2016), each brief generates up to three response options: a *counter-narrative* (direct rebuttal following inoculation principles), an *alternative narrative* (positive reframing addressing the underlying grievance), and a *non-engagement recommendation* when response risks amplification. Each option specifies recommended messenger type, moral frame (Feinberg and Willer, 2015), anti-thesis formulation, and backfire risk (low/medium/high).

Response typology. Drawing on the RAN distinction between counter-narratives and alternative narratives (Radicalisation Awareness Network, 2022; **?**) and the GAMMMA+ framework (Silverman et al., 2016), we require the LLM to generate up to three response options per narrative cluster, each labelled by type:

- **Counter-narrative** (direct rebuttal): truth-first correction of the core false claim, with a cited verifiable source. Follows inoculation principles: name the manipulation technique, then provide the correction (van der Linden et al., 2017).
- **Alternative narrative** (positive reframing): a response that addresses the underlying grievance or emotional need the narrative exploits, without repeating the false claim. Must introduce "something novel" rather than generic values (Radicalisation Awareness Network, 2022).
- **Non-engagement recommendation**: when the narrative is likely to be amplified by any response, or when direct engagement risks reactance, the brief recommends non-engagement with justification (UK Government Communications Service, 2019).

For each response option, the brief must specify: the recommended *messenger type* (e.g., community leader, subject-matter expert, peer voice—not government official, following research showing that peer credibility outperforms institutional authority (**??**)); the *moral frame* aligned with the target

audience's values rather than the communicator's (Feinberg and Willer, 2015); a concrete *anti-thesis formulation* (a model rebuttal sentence or dialogue opening); the *backfire risk* (low/medium/high with justification); and a *confidence level*.

Contextual cards. Crucially, counter-strategies are grounded in domain-specific *contextual cards* containing verified statistics with source URLs, media ecosystem notes, and historical memory triggers. For each narrative cluster's thematic domain (e.g., EU relations, economy, security/defence), the pipeline loads a domain-specific contextual card containing: verified statistics with source URLs (e.g., Eurostat migration data, KSH Hungarian labour market data); media ecosystem notes indicating which outlets can credibly carry a counter-frame, based on RSF Media Ownership Monitor and NMHH data; and historical memory triggers relevant to the narrative. Domain assignment is automatic, with fallback to a general EU-relations card when the domain classifier has low confidence. In the current implementation, contextual cards were initially generated by an LLM and then reviewed by a researcher for factual accuracy—a pragmatic compromise that preserves the architectural constraint (the LLM generating the brief cannot access its own parametric knowledge for facts) while acknowledging that full expert curation would be required for deployment. This directly addresses the well-documented risk that LLMs hallucinate statistics or fabricate sources when generating policy-relevant content (McKenna et al., 2023).

De-escalation principles. The system prompt includes evidence-based principles drawn from inoculation research (van der Linden et al., 2017) and WHO risk communication guidance (World Health Organization, 2017): (1) acknowledge emotion before introducing facts; (2) name the manipulation technique, not the belief; (3) lower certainty rather than flip belief; (4) avoid repetition of the false claim, using truth-first framing; (5) offer agency without mobilisation; (6) know when non-engagement is the safest response. These principles are injected as explicit constraints in the generation prompt, not as suggestions.

Automated validation. Six checks flag briefs for mandatory human review: actor reference (every bullet must reference an input actor), technique hallucination, master narrative consistency, URL provenance (all URLs must originate from the contextual card), source provenance, and statistic traceability.

7. Evaluation

We evaluate at three levels: extraction quality, brief quality (per-section), and failure modes. The central question is which pipeline components can support semi-automation and which require strict human control.

7.1. Extraction Quality

A separate GPT-4o instance evaluates extraction on four dimensions (1–5 Likert): grounding, role accuracy, atomicity, and NPF logic. Evaluation uses 5 randomly sampled articles (`random_state=42`).

Art.	Grnd.	Role	Atom.	NPF
4706	5	4	5	4
264	5	5	4	5
550	5	4	5	4
4326	5	4	5	5
4811	5	4	5	4
Mean	**5.0**	**4.2**	**4.8**	**4.4**

Table 3: LLM-as-judge extraction scores (n=5 articles).

Role accuracy is weakest (4.2), with judge comments noting genuine ambiguity around neutral-vs-slightly-positive portrayals. A translation-first variant (pre-translating Hungarian to English before extraction) produced near-identical quality (all mean deltas $\leq$0.2), suggesting pre-translation does not justify the doubled cost and risk of distortion in culturally specific terminology.

Critical limitation. The judge is the same model family as the extractor. The perfect 5.0 grounding score warrants scrutiny: shared training data may make systematic errors invisible to the judge. On n=5 articles, no statistical claims can be made; we treat these as pilot indicators. For under-resourced languages where qualified annotators are scarce, LLM-as-judge may be the only scalable evaluation, yet it is precisely in these settings that model biases are most concerning.

7.2. Brief-Level Evaluation

We conduct two complementary evaluations: (1) human evaluation with parallel LLM-as-judge on a single brief, and (2) scaled LLM-as-judge across 29 briefs.

Human evaluation (n=4, single brief). Four reviewers independently evaluated Brief C-001 ("Tisza's Contested Characterisation"): a research volunteer (R1), an investigative journalist covering

Eastern European media (R2), an academic specialist in Hungarian politics (R3), and an NGO analyst on democratic governance (R4). Three LLM reviewers (Claude Sonnet, GPT-4o, Gemini Pro) evaluated the same brief using identical forms. Each reviewer assigned Accept/Edit/Reject per section.

	Human (n=4)			LLM (n=3)		
Section	A	E	R	A	E	R
A. Summary	3	1	0	1	2	0
B. Characters	2	2	0	1	2	0
C. Manipulation	2	2	0	3	0	0
D. Escalation	4	0	0	3	0	0
E. Counter-str.	1	2	1	1	2	0
Total	12	7	1	9	6	0

Table 4: Accept (A) / Edit (E) / Reject (R) for Brief C-001.

Scaled LLM-as-judge (n=29). An LLM simulating an experienced Hungarian political journalist evaluated the 29 largest clusters using the same form. These automated assessments cannot verify quotes against Hungarian sources or assess real-world messenger feasibility.

Section	**Accept**	**Edit**	**Reject**
A. Summary	22 (76%)	6 (21%)	1 (3%)
B. Characters	11 (38%)	17 (59%)	1 (3%)
C. Manipulation	5 (17%)	24 (83%)	0
D. Escalation	24 (83%)	5 (17%)	0
E. Counter-str.	10 (34%)	18 (62%)	1 (3%)

Table 5: LLM-as-judge acceptance rates (n=29 briefs).

8. Analysis

8.1. Quality Gradient: Descriptive vs. Prescriptive

The central empirical finding, replicated across both evaluations, is a consistent quality gradient along the descriptive-to-prescriptive axis. Section D (Escalation) achieved unanimous human acceptance and 83% at scale. R4 noted that "the LLM most accurately captured the escalation assessment," suggesting that this section—operating on observable textual signals within a constrained 3-level taxonomy—is amenable to semi-automation. Section A (Summary, 76%) similarly succeeds by staying close to source text.

Section E (Counter-strategy) was weakest: the only Reject in the human evaluation (R2) was motivated by the assessment that recommended strategies would be counterproductive within Hungarian political dynamics. R4 independently noted that suggested messenger types face prohibitive reputational costs. At scale, backfire risk assessment is near-perfect (97% correct), but source veracity is weakest (17% correct), with the contextual card unable to support domestic electoral topics outside its EU/security scope. Actionability: 86% rated "With modification"—usable scaffolding requiring expert adaptation.

Section C's 83% edit rate at scale is dominated by technique over-detection: the system flags standard political rhetoric as manipulation, and master narrative confidence scores show no relationship to analyst-assessed severity.

8.2. Human–LLM Divergence

The overall mean rating difference between human and LLM judges is negligible (+0.05 on a 0–2 scale), but conceals critical section-level divergences. LLMs were more lenient on Section C (+0.50) and E (+0.33). The most consequential asymmetry: **no LLM reviewer issued a Reject for any section**, despite free-text justifications articulating grounds for one. Claude's Section E justification correctly identified that "government media routinely discredits TI data" and non-partisan voices are "increasingly rare"—observations logically supporting a Reject—but rated Edit.

LLM reviewers cluster toward moderate human opinion (Gemini matches R3 at 80%; ChatGPT matches R4 at 80%) but no LLM reproduces R2's critical threshold. Mean inter-LLM agreement (60%) exceeds inter-human agreement (47%), but this reflects lower variance from positivity bias rather than superior evaluation. Notably, LLM justifications sometimes identified problems missed by humans (e.g., ChatGPT flagged attribution blurring), suggesting LLM justifications are more informative than LLM ratings; a finding with implications for hybrid evaluation protocols.

9. Failure Mode Analysis

The scaled evaluation identified 92 failure instances across 29 briefs (mean 3.2/brief). Table 6 presents the taxonomy with counts, pipeline stage, and mitigation level.

Character/role failures (17) divide into: marginal figures inflating character maps (addressable via salience thresholds), abstract entities treated as characters ("forint," "war"—addressable via entity-type filtering), and weaponised-defector roles the schema cannot express (a schema limitation). Missing verified sources (15) follow a clear pattern: the contextual card covers EU/security domains but cannot support domestic electoral counter-strategies. All three Rejects in the scaled evaluation trace to a single brief (C-056: Japanese LDP

Failure mode	n	Stage	Mitigation
Character/role issues	17	Extr.+Gen.	Salience filter; schema
Missing verified sources	15	Gen. (E)	Card expansion
Technique over-detection	9	Extraction	Severity threshold
Cluster overlap	8	Clustering	Similarity checks
Escalation miscalibration	8	Gen. (D)	Cross-brief pass
Analytical conflation	7	Clust.+Extr.	Purity checks
Strategic misalignment †	—	Gen. (E)	Data + human
Messenger mismatch †	—	Gen. (E)	Data-level

Table 6: Failure taxonomy (n=29 briefs; 92 instances). †Identified in human evaluation; not systematically quantifiable by LLM-as-judge at scale.

content merged with Hungarian politics via clustering failure), indicating the generation stage does not independently produce unsafe content from coherent input.

Cross-context narrative collision (C-056). The C-056 cluster combines Hungarian domestic electoral narratives with coverage of the Japanese LDP election. The merging is not due to lexical overlap but to structural similarity at the Narrative Policy Framework level: both contexts contain "leader-as-hero," "electoral victory," and "external validation" (e.g., Trump endorsement) frames. Because clustering operates on abstracted frame representations, it groups narratives with similar role structures even when they belong to entirely separate geopolitical and informational contexts. This reveals a systematic limitation: NPF-level similarity can dominate over contextual coherence, producing clusters that are analytically invalid for downstream use (e.g., counter-strategy generation). Mitigation strategies include: (1) Geopolitical coherence constraints: require clusters to maintain consistent country/entity sets (e.g., Hungary vs. Japan); flag clusters with mixed geopolitical anchors. (2) Entity-distribution checks: enforce that dominant actors (top-k entities) belong to a coherent political system; detect outliers such as "Takaichi Sanae" in a Hungarian cluster. (3) Two-stage clustering: first cluster by context (country/language/domain), then by narrative structure to prevent cross-context collisions.

The human evaluation uniquely identified three failure classes invisible to LLM judges: overconfident intent attribution (flagged by 2/4 humans), unsafe strategy recommendation (1/4), and manipulation over-detection calibration (1/4). The dominant failure across both evaluations is not hallucination but *contextual insufficiency*: outputs are grounded in source text but fail when the task requires knowledge absent from both text and contextual cards.

Automation assessment. Section D is safe for semi-automation with standard review. Sections A–C provide useful scaffolding requiring editorial correction. Section E requires strict expert oversight; its failure severity indicates that the section's value-add over manual drafting is not yet demonstrated.

Labour implications. Human review times ranged from 10 minutes (R1, non-specialist) to 60 minutes (R4, domain expert), correlating with expertise and edit depth. For Sections A and D, where acceptance is high, the brief provides a usable first draft. For Section E, R2 noted that the review burden may exceed manual drafting, because reviewers must identify strategic errors, verify corrections introduce no new backfire risks, and assess political feasibility. This suggests the pipeline's labour savings are section-dependent: genuine for descriptive components, uncertain for prescriptive ones.

10. Limitations

Data and scope. The 574-article corpus is a temporal snapshot; narratives closer to election day (April 2026) are not captured. Sources are RSS news only, excluding social media. The pipeline uses two LLM families (GPT-4o, Claude Sonnet) with no systematic model comparison. Counter-strategies are evaluated for perceived usefulness, not real-world efficacy. Contextual cards were LLM-generated and researcher-reviewed, not fully expert-curated. The response typology simplifies GAMMMA+; we do not implement audience segmentation.

Contextual card reliability. While we did not conduct a formal error-rate audit, researcher review revealed three recurrent error classes: (1) incorrect or non-verifiable source attribution (e.g., statistics linked to secondary rather than primary sources), (2) outdated data (particularly for fast-changing economic indicators), and (3) overly generalised claims lacking country-specific grounding. These errors were typically detectable during manual review and did not propagate into Sections A–D, but they directly affected Section E (counter-strategy grounding). Based on qualitative inspection, source attribution errors were the most frequent, followed by outdated data; fabricated statistics were rare

but cannot be excluded without systematic auditing. This reinforces that contextual cards are currently a bottleneck for safe deployment and require expert curation for high-stakes domains.

Evaluation constraints. Human evaluation covers a single brief with n=4 reviewers—insufficient for statistical inference. The scaled evaluation (n=29) uses LLM-as-judge, which we show underestimates contextually grounded failures and never issues Reject labels. No Hungarian-language annotation benchmark exists for NPF extraction or narrative-level manipulation profiling—a structural challenge for political NLP in under-resourced languages where the scarcity of qualified analysts that motivates automation also limits evaluation. Our manipulation detection was benchmarked only on Russian; cross-linguistic transfer is untested.

11. Ethical Considerations

We scope this work as analyst support, not automated decision-making. No moderation, content removal, or audience targeting decisions are made. We do not perform voter profiling or persuasion optimisation. Counter-strategies are framed as de-escalation with explicit confidence levels and backfire risk assessments. The response typology includes non-engagement recommendations, and de-escalation principles prohibit false-claim repetition.

12. Acknowledgements

The authors sincerely thank I., D. and O. for their evaluation contribution, and M. for help with literature review.

13. Conclusion

We presented a pipeline for generating structured narrative intelligence briefs from Hungarian pre-election news, combining NPF-grounded extraction, SemEval-taxonomy manipulation detection, embedding-based clustering, and constrained counter-strategy generation. The central finding is a consistent quality gradient: escalation assessment (83% accept at scale, unanimous in human evaluation) and narrative summary (76%) succeed because they operate on observable textual signals with constrained output spaces. Counter-strategies (34% at scale; 25% human accept) fail not through hallucination but through contextual insufficiency.

A key open question is whether counter-propaganda frameworks developed for relatively open media environments (GAMMMA+, inoculation, RAN) require adaptation for captured-media contexts where institutional trust and messenger availability are structurally constrained—as both human and LLM reviewers independently observed for the Hungarian case.

An equally important finding concerns evaluation methodology. LLM-as-judge reviewers approximate median human judgment but cannot reproduce the critical tail: no LLM issued a Reject despite articulating grounds for one. This suggests a hybrid critique-then-rate protocol—LLMs generate structured critiques, humans assign severity—as a concrete improvement for evaluating politically sensitive NLP outputs.

From a practitioner perspective, the pipeline compresses $\sim$15 hours of reading into a reviewable format, with genuine labour savings for descriptive-analytical sections. Future work should prioritise expanded human evaluation across multiple briefs, contextual card enrichment with messenger constraints, and field testing of counter-strategies with target populations.

14. Bibliographical References

Graham Brookie and Digital Forensic Research Lab. 2023. Narrative warfare: How the Kremlin and Russian news outlets justified a war of aggression against Ukraine. Technical report, Atlantic Council of the United States.

Carnegie Endowment for International Peace. 2024. Countering disinformation effectively: An evidence-based policy guide. Technical report, Carnegie Endowment for International Peace.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. 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 5636–5646.

Matthew Feinberg and Robb Willer. 2015. From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, 41(12):1665–1681.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762.

Kai Gehring and Matteo Grigoletto. 2023. Analyzing climate change policy narratives with the character-role narrative framework.

Michael D Jones and Mark K McBeth. 2010. A narrative policy framework: Clear enough to be wrong? *Policy studies journal*, 38(2):329–353.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 2758–2774.

Yulia Otmakhova and Lea Frermann. 2025. Narrative media framing in political discourse. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9167–9196, Vienna, Austria. Association for Computational Linguistics.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michał Marcińczuk, Arkadiusz Modzelewski, Ivo Moravski, et al. 2025. Slavicnlp 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 254–275.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th international workshop on semantic evaluation (SemEval-2023)*, pages 2343–2361.

Radicalisation Awareness Network. 2022. Lessons learned from alternative narrative campaigns. Technical report, European Commission, Directorate-General for Migration and Home Affairs.

Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34):eabo6254.

Elizabeth A Shanahan, Michael D Jones, Mark K McBeth, and Claudio M Radaelli. 2018. The narrative policy framework. In *Theories of the policy process*, pages 173–213. Routledge.

Tanya Silverman, Christopher J. Stewart, Zahed Amanullah, and Jonathan Birdwell. 2016. The impact of counter-narratives: Insights from a year-long cross-platform research project. Technical report, Institute for Strategic Dialogue.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

UK Government Communications Service. 2019. RESIST counter-disinformation toolkit. Technical report, UK Government.

Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2):1600008.

World Health Organization. 2017. Communicating risk in public health emergencies: A WHO guideline for emergency risk communication policy and practice. Technical report, Geneva.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

15. Language Resource References

A. Appendix: Prompt Engineering Details

NPF extraction prompt evolution. Version 1: 12-field schema; well-grounded but ∼40% background/logistics pollution. Version 2: expanded to 15 fields including `narrative_function`, `causal_relation_type`; model ignored new fields (prompt signal-to-noise problem). Version 3: added hard background exclusion constraint; improved recall but analytical fields still unpopulated. Version 4: schema-first with inline comments, single worked example, schema reduction to 7+3 fields. Full prompts available in supplementary materials.

Manipulation detection prompt evolution. Version 1: zero-shot custom taxonomy, micro F1=0.008 (1 TP, 6 FP, 235 FN). Version 2: SemEval labels + 4 examples, F1=0.104 (20 TP, 130 FP, 216 FN); systematic failures: phantom labels (24 FP), span mismatch (17 near-misses), zero recall on 12/25 categories, multi-label blindness (51% of gold spans were multi-label). Version 3 (production): 12 language-specific examples, 10-row disambiguation table, explicit multi-label instruction, hard vocabulary constraint; F1=0.227 (45 TP, 115 FP, 191 FN). Adapted for Hungarian production with domain-specific

examples and removed competition-specific calibrations.

B. Appendix: BERTopic Comparison

Five BERTopic configurations were compared using `paraphrase-multilingual-mpnet-base-v2` embeddings. Default parameters produced degenerate clusterings (one mega-cluster); with tuned parameters, raw text produced ∼46 event-level topics (silhouette 0.71, C_v=0.60) and NPF summaries produced ∼38 narrative-adjacent topics (silhouette 0.66, C_v=0.62). The fundamental finding: BERTopic recovers *topics* but not *narratives*—it lacks actor–role assignments and causal logic.

C. Appendix: Manipulation Taxonomy

The SemEval/SlavicNLP taxonomy comprises 6 categories and 25 labels: Attack on Reputation (name_calling, guilt_by_association, doubt, appeal_to_hypocrisy, questioning_reputation), Justification (flag_waving, appeal_to_authority, bandwagon, appeal_to_fear, appeal_to_values), Distraction (straw_man, whataboutism, red_herring, appeal_to_pity), Simplification (causal_oversimplification, false_dilemma, consequential_oversimplification, false_equivalence), Call (slogan, conversation_killer, appeal_to_time), Manipulative Wording (loaded_language, obfuscation, exaggeration_minimisation, repetition). For brief generation, these are remapped to 6 intent-based categories: legitimisation, emotional manipulation, identity polarisation, simplification, distraction, and mobilisation.

D. Appendix: Prompts

D.1. NPF-specific Summarisation

System Role:

```
You are an expert in the Narrative
    Policy Framework (NPF).
```

User Task Prompt:

```
Analyze the following Hungarian news
    text.
    Your goal is to rewrite the content
    focusing strictly on narrative
    elements.

    Extraction Rules:
    1. Identify the SETTING
    (context/environment).
    2. Identify CHARACTERS: Hero
    (problem solver), Villain (cause of
    problem), Victim (harmed party).
    3. Identify the PLOT: Is it a story
    of Decline, Stalled Progress, or
    Illusion of Control?
    4. Identify the MORAL: What is the
    proposed solution or lesson?

    Constraints:
    - If there is NO clear narrative
    (e.g., pure factual reporting),
    output a factual summary only.
    - If the MORAL is ambiguous or
    missing, omit it. DO NOT
    hallucinate a lesson.
    - If there are MULTIPLE Viewpoints,
    explicitly state both narrative
    arcs.

    Output Format:
    Return a coherent paragraph in
    Hungarian.

    Text:
    {{article}}

    Summary:
```

Model Configuration:

```
model: gpt-4o
temperature: 0
```

D.2. Manipulation Techniques Extraction

System Role:

```
ROLE: You are a specialist in detecting
    propagandistic manipulation
    techniques in political texts,
    following the predefined persuasion
    technique taxonomy.

TASK: Extract ALL instances of
    manipulation techniques from the
    input text. For each instance,
    identify:
1. The technique label (from the
    controlled vocabulary below)
2. The verbatim text span where the
    technique operates

=== FEW-SHOT EXAMPLES ===
Each example shows the FULL PASSAGE to
    extract.
Multiple techniques often apply to the
    SAME passage  output each
    separately.

EXAMPLE 1  causal_oversimplification +
    loaded_language (MULTI-LABEL on
    same span):
Passage:
"Trump most egyrtelmen kirly a
    sakktbln. gy jtszotta meg a partit,
```

```
    hogy Zelenszkij brmely dntse az
    USA-nak kedvez. Ha folytatdik a
    hbor, Oroszorszg tovbb gyengl. Ha
    vge, Ukrajna gazdasgi bekebelezse
    kvetkezik. Minden forgatknyv
    profit."
 causal_oversimplification: Bonyolult
    geopolitikai folyamatokat egyetlen
    szerepl okos hzsaira" redukl;
    egyetlen ok-okozati lnc felttelezse.
 loaded_language: kirly", jtszotta meg
    a partit", profit"  rzelmileg
    tlfttt, nem semleges megfogalmazs.
Note: The ENTIRE multi-sentence passage
    is one unit. Do NOT split into
    sentence-level spans.

EXAMPLE 2
    consequential_oversimplification +
    appeal_to_fear (MULTI-LABEL):
Passage:
"Magyarorszg slyos demogrfiai vlsgban
    van. Minden vlsg, ha egyszer nem
    trjk meg, nem lltjuk meg, vgl
    katasztrfhoz vezet. Egy nemzet
    kihalsa lass folyamat, vtizedekig
    tarthat, s kzben megszokjk, nem
    veszik szre. A vgn belenyugszanak,
    s felhagynak az ellenllssal."
 consequential_oversimplification:
    Elkerlhetetlen katasztrft llt (vlsg
     kihals) rnyals nlkl, csszs lejt
    logikval.
 appeal_to_fear: A nemzeti kihals
    flelmvel mobilizl.

EXAMPLE 3  obfuscation:
Passage:
"Vita folyik arrl, hogy mindez a
    nyugati polgri forradalmak
    kvetkezmnye-e, amelyek felszmoltk a
    monarchit s az egyhzat, s az
    Aranyborj imdatt hoztk el. Vajon a
    valls befolysolja-e a csaldi
    rtkeket, amelyeket a vrosi letmd
    rombolt le? A tbbgyermekes csald
    paraszti racionalits volt? s
    mindezt a kapitalizmus puszttotta
    el?"
 obfuscation: Egymsra halmozott
    retorikai krdsek, vilgos llts nlkl;
    elre sugallt, de ki nem mondott
    kvetkeztets.

EXAMPLE 4  straw_man +
    name_calling_labelling
    (MULTI-LABEL):
Passage:
"Zelenszkij, akinek pszichjt lltlag
    folyamatos droghasznlat roncsolta,
    id eltt elhagyta a Fehr Hzat. Trump
    ezt kveten kijelentette, hogy
    Zelenszkij nem ksz a bkre, s
    tiszteletlen volt az Egyeslt
    llamokkal szemben. Ha ez nem
    sznjtk, akkor Trump lnyegben szabad
    kezet ad Putyinnak."
 straw_man: Az ellenfl llspontjnak
    karikatraszer torztsa (instabil,
    drogos).
 name_calling_labelling: pszichjt
    droghasznlat roncsolta"  megblyegz
    cmkzs hiteltelents cljbl.

EXAMPLE 5  whataboutism:
Passage:
"Az interj nagy visszhangot vltott ki,
    de az Egyeslt llamokban azonnal egy
    lltlagos orosz nukleris mholdrl
    kezdtek beszlni, amelyrl ksbb
    kiderlt, hogy lhr. Mr az is
    sokatmond, hogy ezzel prbltk
    elterelni a figyelmet."
 whataboutism: Az rdemi krds helyett
    msik fl lltlagos manipulcijra
    terels.

EXAMPLE 6  appeal_to_time:
Passage:
"Egyre inkbb az az rzsem, hogy versenyt
    futunk az idvel."
 appeal_to_time: Az id srgetse nmagban
    vlik rvv.

EXAMPLE 7  exaggeration_minimisation +
    slogan:
Passage:
"2004 ta Magyarorszg npessge akr meg is
    duplzdhatott volna. Bks vek voltak,
    vlsg nlkl, a felemelkeds korszaka,
    fellls a trdrl". Mirt nem trtnt
    meg? Nem volt pnz? Dehogynem. vente
    tzmillirdok tntek el."
 exaggeration_minimisation: A
    megduplzdhatott volna" ersen tlz
    llts.
 slogan: fellls a trdrl"  rzelmi
    rvidts, politikai jelsz.

EXAMPLE 8  guilt_by_association:
Passage:
"Ezek a bankovai neoncik s tmogatik
    ugyanazon az ton jrnak, mint
    egykori pldakpeik a Harmadik
    Birodalomban, a Hitlerjugenddel."
 guilt_by_association: Az ellenfl
    sszekapcsolsa ncikkal, erklcsi
    megblyegzs cljbl.

EXAMPLE 9  red_herring:
Passage:
A szveg a demogrfiai problmkrl indul,
    majd hirtelen a Rmai Birodalom,
    Biznc, az aztkok s Atlantisz buksrl
    kezd beszlni, vgl Magyarorszg
    eltnst vetti elre.
```

```
 red_herring: A konkrt szakpolitikai
    krdsrl rzelmi-historikus
    prhuzamokra val elterels.
Also: consequential_oversimplification
    a hanyatls  sszeomls lnc
    leegyszerstse.

EXAMPLE 10  conversation_killer:
Passage:
"Minden a
    tagadsharagalkudozsdepresszielfogads
    smja szerint zajlik. Most valahol a
    harag s a depresszi kztt vagyunk.
    Kemny? Lehet. De igaz."
 conversation_killer: Kemny? Lehet. De
    igaz."  a vita lezrsa,
    megkrdjelezhetetlensg sugallata.

EXAMPLE 11  appeal_to_values:
Passage:
"A Nyugat vszzadok ta a felszabadts
    jelszavval rombolja le az emberi
    kzssgek alapjait: az egyhzat, a
    csaldot, a hagyomnyt, a hitet, s
    helykbe az individualizmust s
    erklcsi relativizmust lltja."
 appeal_to_values: Hagyomny, valls,
    csald, erklcs mint pozitv rtkek
    mozgstsa.

EXAMPLE 12  bandwagon
    (appeal_to_popularity):
Pattern:
Ms orszgok mr rgen felismertk", Az egsz
    vilg ltja", Mindenki tudja, hogy"
 bandwagon: llspont igazolsa vlt tbbsgi
    egyetrtssel.

=== CRITICAL RULES ===

RULE 1  SPAN LENGTH:
- If a technique pervades a paragraph,
    quote the ENTIRE paragraph.

RULE 2  MULTI-LABEL DETECTION:
The SAME passage can have 25 techniques
    simultaneously.
- Output SEPARATE JSON entries for each
    technique, even if the evidence
    span is identical.

RULE 3  DISAMBIGUATION TABLE (commonly
    confused labels):
| If tempted to use... | Use instead...
    | Key distinction |
|---|---|---|
| "scapegoating" |
    guilt_by_association,
    appeal_to_fear, or
    name_calling_labelling |
    "Scapegoating" is NOT in this
    taxonomy. Use guilt_by_association
    when linking opponent to a
    negatively perceived group;
    appeal_to_fear when exploiting
    prejudice; name_calling_labelling
    for pejorative labels. |
| "appeal_to_tradition" |
    appeal_to_values | appeal_to_values
    covers tradition, religion,
    morality, heritage. No separate
    tradition label exists. |
| "dehumanisation" |
    consequential_oversimplification or
    name_calling_labelling | Use
    consequential_oversimplification
    for "this will lead to
    dehumanization" arguments;
    name_calling_labelling for
    dehumanizing labels. |
| "black_and_white_fallacy" |
    false_dilemma | Same concept, use
    false_dilemma. |
| "appeal_to_anger" | appeal_to_fear or
    loaded_language | Not in taxonomy.
    Fear/prejudice exploitation
    appeal_to_fear. Emotionally charged
    wording  loaded_language. |
| appeal_to_fear for "X caused Y" logic
    | causal_oversimplification | If
    the text attributes a complex
    outcome to a single cause, that's
    causal_oversimplification  even if
    the tone is fearful. appeal_to_fear
    = leveraging fear/prejudice as the
    primary mechanism. |
| doubt for vague, unclear text |
    obfuscation | doubt = questioning
    credibility/trustworthiness of an
    entity. obfuscation = deliberately
    unclear, vague, or confusing
    language/reasoning. |
| loaded_language for a label/insult |
    name_calling_labelling | If a
    specific person/group receives a
    pejorative LABEL ("fascists", "drug
    addict", "puppets"), that's
    name_calling_labelling.
    loaded_language = emotionally
    charged WORDING that doesn't
    function as a direct label for an
    entity. |
| questioning_reputation for citing
    authority | appeal_to_authority |
    questioning_reputation = ATTACKING
    someone's reputation/character.
    appeal_to_authority = CITING
    someone/something as credible to
    justify a claim. |
| appeal_to_fear for "if A then B then
    disaster" |
    consequential_oversimplification |
    Slippery slope chains (if A  then B
     then C catastrophe) =
    consequential_oversimplification. |
```

```
RULE 4  LABEL VOCABULARY:
Use ONLY these 25 labels. Any other
    label is INVALID:
["appeal_to_authority",
    "appeal_to_fear",
    "appeal_to_hypocrisy",
    "appeal_to_pity", "appeal_to_time",
    "appeal_to_values", "bandwagon",
    "causal_oversimplification",
    "consequential_oversimplification",
    "conversation_killer", "doubt",
    "exaggeration_minimisation",
    "false_dilemma",
    "false_equivalence", "flag_waving",
    "guilt_by_association",
    "loaded_language",
    "name_calling_labelling",
    "obfuscation",
    "questioning_reputation",
    "red_herring", "repetition",
    "slogan", "straw_man",
    "whataboutism"]

=== OUTPUT FORMAT ===
Return ONLY valid JSON:
{
  "text_id": "<string>",
  "techniques": [
    {
      "technique": "<label from
    vocabulary>",
      "evidence_span": {
        "text": "<verbatim quote from
    source, typically 150600 chars>"
      },
      "confidence": "high | medium |
    low"
    }
  ]
}
```

Model Configuration:

```
model: gpt-4o
temperature: 0
```

D.3. NPF Judge

System Role:

```
Your task is to audit the quality of a
    JSON narrative extraction performed
    on a Hungarian news article.

# INPUT DATA
1. Original Article (Hungarian)
2. Extracted Narrative Elements (JSON)

# EVALUATION RUBRIC (Score 1-5)
1. Grounding (Are quotes real?):
   - 5: All 'evidence_span' texts
    appear verbatim or near-verbatim in
    the source.
   - 1: Fabricated quotes or
    hallucinations.

2. Role Accuracy (Hero/Villain/Victim):
   - 5: The 'portrayed_role' perfectly
    matches how the text constructs the
    actor (e.g., if the text praises X,
    X is not a villain).
   - 1: Major misclassification of
    roles (e.g., calling the attacker a
    'victim' without textual evidence).

3. Atomicity (One Claim Per Element):
   - 5: Each element contains exactly
    one distinct claim/action.
   - 1: Multiple unrelated claims are
    merged into a single element.

4. NPF Logic (Causality & Function):
   - 5: Causal claims (if present) use
    correct markers. Narrative
    functions (mobilisation vs claim)
    are correctly identified.
   - 1: Random assignment of narrative
    functions.

# OUTPUT FORMAT
Return valid JSON only:
{
  "score_grounding": <int 1-5>,
  "score_role_accuracy": <int 1-5>,
  "score_atomicity": <int 1-5>,
  "score_npf_logic": <int 1-5>,
  "reasoning": "<Concise explanation of
    the score, citing specific errors
    if any>",
  "missing_major_elements": "<List any
    MAJOR actors or claims that were
    completely missed, or 'None'>"
}
```

Model Configuration:

```
model: gpt-4o
temperature: 0
```

D.4. Narrative Frame Extraction

System Role:

```
You are performing narrative analysis
    grounded in the Narrative Policy
    Framework (NPF).
Your task is to extract atomic
    narrative elements from a Hungarian
    news article.
These elements are analytical
    abstractions for narrative
    intelligence and escalation
    analysis,
not factual event logs and not moral or
    legal judgments.
```

User Task Prompt:

```
# TASK
Extract atomic narrative elements from
    the article below. Follow the
    schema exactly. Every field is
    required  use null where
    appropriate but never omit a field.

# SCHEMA (all fields mandatory)
'''json
{
  "narrative_elements": [
    {
      "actor": "exact surface form as
    in text",
      "actor_type": "individual |
    institution | group | abstract |
    media | unspecified",

      "speaker_role": "hero_narrator |
    narrator | neutral | unclear",
      // Who is the speaker performing
    as? If mobilising, defending, or
    positioning as protector
    hero_narrator. If reporting
    narrator. If ambiguous  unclear.

      "portrayed_role": "hero | villain
    | victim | narrator | neutral |
    unclear",
      // How does the TEXT construct
    this actor? Not your judgment. Must
    be supported by evidence_span.

      "narrative_function": "claim |
    evaluation | threat | mobilisation
    | warning | policy_promise",
      // MANDATORY. Choose ONE. See
    decision table below. Do NOT use
    background or logistics  those
    should not be extracted.

      "action_or_claim": "what the
    actor does, says, or is described
    as doing",

      "target": "exact surface form, or
    null",
      "target_type": "individual |
    institution | group | abstract |
    unspecified | null",
      "portrayed_target_role": "hero |
    villain | victim | threat | proxy |
    neutral | unclear | null",
      // threat = trigger entity that
    causes catastrophe; proxy = stands
    in for a larger villain

      "causal_claim": {
        "text": "explicit causal
    statement from the text, or null",
        "causal_marker_surface":
    "because | therefore | if | when |
    so_that | implied | null",
        // Use the actual logical
    connector. Not grammatical
    particles (hogy, onnan, ami are NOT
    causal markers unless they
    introduce a cause-effect).
        "causal_relation_type": "threat
    | prevention | retaliation |
    inevitability | zero_sum |
    justification | null"
        // What kind of causal logic?
    threat = X causes harm; prevention
    = X stops harm; justification = X
    explains Y.
      },

      "normative_direction": "explicit
    demand, prohibition, or call  or
    null",
      "normative_type": "prohibition |
    obligation | encouragement |
    warning | prediction | null",
      // prohibition = must NOT;
    obligation = must; encouragement =
    should/rally; warning = if not,
    then harm; prediction = will happen

      "constructed_audience": "explicit
    in-group label from the text (e.g.
    'mi', 'magyarok', 'fiatalok'), or
    null",
      // Only if the text explicitly
    constructs a 'we' or addresses a
    defined group.

      "attribution": "direct_quote |
    attributed_claim |
    journalistic_paraphrase |
    implicit_frame",
      "evidence_span": "short quoted
    fragment in original Hungarian",
      "confidence": "high | medium |
    low"
      // high = explicit, central,
    unambiguous. medium = clear but
    secondary or compressed. low =
    implied or structurally incomplete.
    }
  ]
}
'''

# DECISION TABLE: narrative_function
| If the element... |  assign |
|---|---|
| Asserts something as fact or history
    | claim |
| Judges an actor, event, or state as
    good/bad/unfair | evaluation |
| Frames conditional or actual harm
    ("if X then catastrophe") | threat |
| Calls for resistance, action,
    opposition, or rally | mobilisation
    |
```

```
| Frames conditional or preventative
    danger ("beware of X") | warning |
| Commits to a future action,
    programme, or pledge |
    policy_promise |
| Is purely logistical (travel,
    scheduling) or background (context,
    history) with no narrative pressure
    | **DO NOT EXTRACT** |

# DECISION TABLE: confidence
| Condition |  assign |
|---|---|
| Central claim, direct quote,
    unambiguous framing | high |
| Secondary point, condensed
    paraphrase, or one of several
    merged claims | medium |
| Implied, structurally incomplete, or
    analytically debatable | low |

# WORKED EXAMPLE
Input fragment: A magyar kormny ellen
    nem nagy kunszt lzadni, lzadjatok
    Brsszel ellen, onnan fenyegetnek
    minket"  zente a fiataloknak a
    kormnyf.

Correct extraction:
'''json
{
  "actor": "Orbn Viktor",
  "actor_type": "individual",
  "speaker_role": "hero_narrator",
  "portrayed_role": "hero",
  "narrative_function": "mobilisation",
  "action_or_claim": "lzadjatok Brsszel
    ellen, onnan fenyegetnek minket",
  "target": "Brsszel",
  "target_type": "institution",
  "portrayed_target_role": "villain",
  "causal_claim": {
    "text": "onnan fenyegetnek minket",
    "causal_marker_surface": "implied",
    "causal_relation_type": "threat"
  },
  "normative_direction": "lzadjatok
    Brsszel ellen",
  "normative_type": "encouragement",
  "constructed_audience": "fiatalok",
  "attribution": "direct_quote",
  "evidence_span": "lzadjatok Brsszel
    ellen, onnan fenyegetnek minket",
  "confidence": "high"
}
'''

Why this is correct:
- speaker_role = hero_narrator because
    Orbn positions himself as rallying
    protector
- narrative_function = mobilisation
    because it is a direct call to act
- portrayed_target_role = villain
    because Brsszel is framed as the
    source of threat
- causal_relation_type = threat because
    the causal logic is "they threaten
    us"
- normative_type = encouragement
    because it's a rallying imperative,
    not a prohibition
- constructed_audience = fiatalok
    because the text explicitly says
    "zente a fiataloknak"

# RULES (compact)
1. Extract only what is stated or
    linguistically signalled. No
    inferred intent, ideology, or
    hidden strategy.
2. One element = one actor + one
    claim/evaluation/threat/call. Do
    not merge unrelated claims.
3. Use exact surface forms for actors
    and targets. Do not normalise or
    translate names.
4. Do NOT extract background or
    logistics that carry no narrative
    pressure (historical recall, travel
    plans, procedural facts). If in
    doubt, omit.
5. Assign portrayed_role only where the
    text constructs it. If a speaker
    mobilises, defends, or protects
    hero or hero_narrator for
    speaker_role. If weak or mixed
    unclear.
6. Populate causal_claim only for
    explicit causation. Grammatical
    connectors (hogy, ami, amelybl) are
    NOT causal markers unless they
    introduce causeeffect logic. If
    causality is implied but weak, use
    causal_marker_surface = "implied"
    and confidence = "low".
7. Populate normative_direction only
    for explicit demands, prohibitions,
    obligations, or calls. "Therefore
    we must" is only valid if
    explicitly stated.
8. Do NOT infer hostility from
    criticism. Do NOT assign
    hero/villain to journalists. Do NOT
    invent policy goals.

# OUTPUT
Valid JSON only. Hungarian for text
    fields. English for labels. No
    commentary outside JSON.

Article:
{{article}}
```

Model Configuration:

```
model: gpt-4o
```

```
temperature: 0
```