Optimizing LLM-based Discourse Analysis in Low-Resource Languages: A Case Study of Bulgarian Euro Adoption Narratives

Dobromir Tsolyov
GATE Institute / Sofia, Bulgaria
dobromir.tsolyov@gate-ai.eu

Keith Kiely
GATE Institute / Sofia, Bulgaria
keith.kiely@gate-ai.eu

Georgi Goranov
GATE Institute / Sofia, Bulgaria
georgi.goranov@gate-ai.eu

Iglika Ivanova
GATE Institute / Sofia, Bulgaria
iglika.ivanova@gate-ai.eu

Stanimir Dimitrov
GATE Institute / Sofia, Bulgaria
stanimir.dimitrov@gate-ai.eu

Abstract

This study explores the "complexity ceiling" of LLMs in analyzing culturally specific Euro adoption narratives in Bulgarian. Navigating this high-subtext content requires pipelines that balance abstract reasoning with rigorous stability. Using the IWH and FABLE methodologies, which comprise a total of 11 dimensions scored $\{0, 1, 2\}$, we evaluated model performance on a dataset sampled for diversity via vector embeddings.

We operationalize a complexity ceiling as the highest-complexity configuration within a bounded search space that satisfies a predefined stability criterion. The ceiling was identified via "descent to feasibility": beginning from the highest complexity parameters within the predefined search space, and iteratively reducing internal and external parameters until the stability criterion was met. Stability was measured using the "80/80 rule," requiring at least 80% of samples to maintain modal consistency in 8 out of 10 independent iterations, across all dimensions. Results indicate that pipelines capable of both stability and higher-order reasoning are achievable through calibrated trade-offs between internal and external complexity. The leading configuration met the 80/80 criteria for 10 out of 11 dimensions, suggesting that stable multi-dimensional discourse analysis is achievable under carefully calibrated complexity conditions, even in low-resource settings.

Keywords: LLMs, low-resource languages, critical discourse analysis, political speech

1 Introduction

Automated discourse analysis of politically charged content in low-resource languages presents a distinctive class of challenges that general LLM benchmarks are ill-equipped to anticipate. Bulgarian political narratives surrounding the Euro adoption debate are dense with cultural subtext, historical analogy, and idiomatic framing that resist straightforward analytical decomposition. Recent evaluations in other less resourced European languages such as Latvian and Polish similarly report degraded performance and heightened vulnerability in multilingual LLMs, supporting the view that low-resource politico-cultural contexts remain a structurally hard regime for current models (Dargis et al., 2024). When an LLM is tasked with scoring such content across multiple abstract dimensions simultaneously, the interaction between linguistic complexity and inferential load creates conditions under which output consistency degrades — not uniformly across dimensions, but in ways that are specific to the combination of task structure, prompt design, and model configuration. Standard performance benchmarks, calibrated on cleaner, higher-resource tasks, offer no reliable basis for predicting where this degradation begins or how severe it will be (Gürses and Ceylan, 2026).

This unpredictability is the core problem this study addresses. Applying the IWH and FABLE frameworks (Kiely, 2025; Patel et al., 2025; Sehat et al., 2024) to Bulgarian Facebook posts requires a pipeline that is not merely capable of higher-order

reasoning, but demonstrably stable under repeated inference to be trusted as an analytical instrument. Without this stability, scores cannot be meaningfully interpreted as evidence of discourse properties, regardless of how sophisticated the underlying model is. Inversely, stability without sufficient reasoning capacity would lead to consistently inaccurate results. The research objective is therefore to identify high-performing pipeline configurations that jointly balance scoring stability and reasoning depth within politically and culturally complex content.

Unique contributions: We contribute (1) an operational definition and search procedure for an LLM complexity ceiling on high-subtext political discourse in a low-resource linguistic environment, (2) a stability benchmark (80/80) tailored to multi-dimensional discourse scoring, and (3) empirical evidence about how internal and external complexity interact in a low-resource, real-world setting.

2 Theoretical Framework

The theoretical framework for this study integrates geopolitical power cycle theory with multi-stage misinformation harm assessments to analyze automated discourse in low-resource linguistic environments. By combining the **Indicators of Weak Hegemony (IWH)** with the **FABLE** framework, this study evaluates the complexity ceiling of Large Language Models (LLMs) in navigating high-subtext sociopolitical narratives.

2.1 Literature Review: Social Science Approaches to Hegemony and Media Deception

Hegemonic power is commonly understood as cyclical, moving through phases of rise, consolidation, and decline under pressures such as military overreach, institutional erosion, and domestic polarization (Patel et al., 2025: 10). In a Gramscian tradition, hegemony refers not only to coercive capacity but to the ability of dominant groups to manufacture consent through culture, political institutions, and everyday social relations (Kiely, 2025: 2). Within this cycle, a weak hegemony is marked by the fragmentation of once-dominant narratives and the erosion of trust in democratic institutions, as competing projects struggle to redefine what counts as “common sense” (Kiely, 2025: 1). Recent work on LLMs’ political worldviews shows that model outputs on ideological questions can vary systematically across prompts, and that reliability and cross-prompt consistency are themselves important objects of study in political communication research (Ceron et al., 2024; Alali et al., 2026) The Indicators of Weak Hegemony (IWH) translate this abstract trajectory into observable dimensions: Fragmented Consensus captures discursive breakdowns of liberal or democratic consensus, Counter-Hegemonic Forces track the visibility of actors and narratives that challenge the established order, and Crisis of Authority reflects contestation of institutional legitimacy and moral-intellectual leadership (Patel et al., 2025: pp. 8–9; Kiely, 2025: p. 3). These and related indicators, such as increased reliance on coercion over consent, are encoded as separate analytical dimensions in our annotation scheme.

Resistance to hegemonic decline is mirrored in the information ecosystem, where media deception exploits cognitive heuristics—rules of thumb—to shape public perception with minimal argumentative substance (Olariu, 2022: 30). Research in political psychology and media studies identifies analytical thinking as a key dynamic shield against such manipulation (Olariu, 2022: 29). In this context, implicit claims function as assertions conveyed through rhetorical strategies or contextual gaps rather than explicit factual statements, steering interpretation without direct evidential support (Nair, 2025: 27). These mechanisms of deception and inference are central to the FABLE dimensions discussed below and motivate our focus on how reliably an LLM can score such high-subtext content across multiple socio-political indicators.

2.2 Indicators of Weak Hegemony (IWH)

The IWH framework operationalizes structural markers of decline to monitor socio-political stability Patel et al. (2025: 8). It defines three primary dimensions:

Fragmented Consensus: This measures the breakdown of a unified public or political agreement. Historically, internal division and the death of a "liberal consensus" signal an irreversible phase of institutional decay (Patel et al., 2025: 9).

Counter-Hegemonic Forces: This assesses the presence of narratives or groups that actively challenge the established status quo. (Such forces often include "organic intellectuals" emerging from marginalized groups to advance alternative worldviews (Kiely, 2025: 3).

Crisis of Authority: This evaluates the perceived loss of legitimacy in established institutions and leadership. It is often exacerbated when leadership styles contribute to the erosion of moral authority, appearing to place figures above the law (Patel et al., 2025: 70).

In the broader IWH scheme, these sit alongside additional indicators such as reliance on coercion rather than consent and deficits of moral-intellectual leadership, which we encode as separate analytical dimensions in our annotation scheme.

2.3 The FABLE Framework: Measuring Misinformation Harm

To prioritize fact-checking efforts, this study utilizes the FABLE framework, which assesses the potential urgency and magnitude of misinformation as a harm (Sehat et al., 2024: 310). Unlike binary truth-matching, FABLE evaluates content across five variable dimensions (Sehat et al., 2024: 373):

Fragmentation (Social): The degree to which information breaks down social cohesion or questions the trust and functioning of public institutions (Sehat et al., 2024: 376).

Actionability: The extent to which content encourages the audience to take specific, often disruptive, actions, with an emphasis on potential physical harm (Sehat et al., 2024: 377).

Believability: The effectiveness of a message to a specific community, often measured by the absence of high-quality, publicly accessible refutations (Sehat et al., 2024: 380).

Likelihood of Spread: The potential for virality, determined by the "influencer" status of the spreader or the content's format, such as sensationalist text or video (Sehat et al., 2024: 381–382).

Exploitativeness: The degree to which the narrative preys on human fears, emotional triggers, or the lack of resources in vulnerable groups like children or the elderly (Sehat et al., 2024: 383–384).

2.4 Operationalizing Believability: The Role of Logical Fallacies

The dimension of **Believability** (FABLE-3) is operationalized through the identification of **logical fallacies**, which serve as rhetorical "fingerprints" used to counterfeit argumentation (Olariu, 2022: 33). These fallacies act as markers that artificially increase content credibility by bending cognitive heuristics (Olariu, 2022: 30). Key fallacies include:

Anecdotal Evidence (The Anecdote): Using a particular situation to generalize and ascribe general significance or representation power (Olariu, 2022: 33). This generalizes personal stories to broader worldviews that shape understanding and behavior Nair (2025: 17).

Slippery Slope: A fallacy used as a "prophylactic schemata" to avoid perceived emotional costs by arguing that a small first step will lead to catastrophic events (Olariu, 2022: 31).

In practical application, identifying these tactics allows LLMs to decompose content into rhetorical strategies (Jeong et al., 2025: 6934). However, Believability has proven to be a "problematic dimension" for automated analysis, frequently failing to meet the 80/80 standard.

3 Methodology

3.1 Data Sampling and Scoring

3.1.1 Data Sampling

The data used in these experiments contains Facebook posts from 02.05.2025 to 22.06.2025, centered around the debate around former president Rumen Radev's proposal for a referendum on joining the Eurozone in 2026. It was obtained through the Meta Content Library, filtered by the keywords "euro" and "referendum". It initially consisted of 445 Facebook posts from various public profiles. For this study, we analyzed only the textual data and ignored any other media. Many exact duplicates were removed as well, with most of them being the same post, shared from multiple profiles. The cleaned dataset we took our samples from consists of 302 posts.

For the optimal pipeline search, the objective was to identify a subset that was both sufficiently representative of the full dataset's diversity and small enough to make 10-run stability testing across all candidate configurations computationally feasible within the available budget. Therefore, 7% of the data was sampled in a way to ensure maximum diversity. First, all of the posts were separated into three groups, based on their previously annotated euro adoption stance ("pro", "anti" and "neutral"). For each group, the data was vectorized using the BAAI/bge-m3 model from HuggingFace Chen et al. (2024) for embedding, and separated into clusters using KMeans. The sample closest to each centroid

was chosen for a test sample. To avoid bias, we selected an equal number of representatives from each group.

3.1.2 Scoring

Political scientists have mapped the theoretical foundations to the specifics of the analysis by translating them into a total of 62 questions in English with possible answers: “Yes”, “No”, and “Unclear” (Appendix E).

The 62 questions were derived by operationalizing abstract theoretical dimensions, based primarily on the work of Gramsci (1971), referred to as the IWH and FABLE frameworks into observable binary indicators specifically tailored to the Bulgarian Eurozone debate. Political scientists on our team mapped each of the 11 dimensions — six from IWH (e.g., Fragmented Consensus) and five from FABLE (e.g., Believability) — to a set of 4–15 questions designed to detect specific rhetorical strategies or sociopolitical markers. For example, under IWH-1 (Fragmented Consensus), Q2 ("Does it contrast specific demographic groups?") was constructed to capture the "two societies" narrative central to Bulgarian polarization. Similarly, FABLE-3 (Believability) is operationalized through questions like Q37 ("Are personal anecdotes used as representative?"), which identify logical fallacies that serve as "rhetorical fingerprints" to artificially increase content credibility. Each question carries a weight reflecting its theoretical importance and empirical power to discriminate between narrative types in the Bulgarian context.

Question weights were established based on the core principle of theoretical importance and empirical discriminating power within the Bulgarian socio-political context. We assigned weights between -1 and 1 to each question to capture both confirmatory and disconfirmatory evidence. For example, in IWH-1 (Fragmented Consensus), Q2 regarding "Us/Them contrasts" is assigned a higher weight (0.6) than Q5 (0.2) because it is considered a more central indicator of the specific "two societies" narrative in Bulgaria.

The system features a bidirectional weighting approach, most notably in the FABLE-3 (Believability) dimension: Positive weights (+0.2 to +0.4): Applied to "high-susceptibility" fallacies like anecdotal evidence or slippery slopes, which empirical research (Marin et al., 2024) suggests are harder for audiences to detect and thus increase a message’s believability. Negative weights (-0.1 to -0.3): Applied to "low-susceptibility" fallacies like ad hominem attacks or extreme polarization, which audiences can often recognize as flawed, thereby decreasing the content’s perceived credibility. Ultimately, these weighted "Yes" responses are summed to produce a raw dimension score, which is then mapped to a final ordinal score (0, 1, 2) through dimension-specific thresholds designed to attenuate noise and ensure analytical stability.

Let the full question set be $Q = q_1, q_2, ..., q_{62}$, where each question q_k belongs to exactly one dimension $d \in D$ and one methodology $m \in \{IWH, FABLE\}$, while every dimension is associated with between 4 and 15 questions. Each question q_k has an associated weight $w_k \in [-1, 1]$, capturing both confirmatory and disconfirmatory evidence. Each question receives a response $a_{i,k} \in \{Yes, No, Unclear\}$ for text i. Only "Yes" contributes to the score:

$$y_{i,k} = 1; \quad (\text{if } a_{i,k} = \text{Yes})$$
$$y_{i,k} = 0; \quad (\text{if } a_{i,k} = \text{No, Unclear})$$

Let $Q_d \subseteq Q$ denote the subset of questions belonging to dimension d. The raw score for text i on dimension d is: $r_{i,d} = \sum_{k \in Q_d} w_k \cdot y_{i,k}$ Each dimension d has a set of dimension-specific thresholds $\theta_{d,1} < \theta_{d,2}$ that map the raw score to the final ordinal score $f_{i,d} \in \{0, 1, 2\}$:

$$f_{i,d} = \begin{cases} 0, & \text{if } r_{i,d} < \theta_{d,1} \\ 1, & \text{if } \theta_{d,1} \leq r_{i,d} < \theta_{d,2} \\ 2, & \text{if } r_{i,d} \geq \theta_{d,2} \end{cases}$$

This elaborate score system thus plays the role of a first-order stability measure, as it attenuates output variance from higher-order reasoning by introducing a discrete buffer against noise and minor fluctuations.

The final output of the LLM pipeline consists of three components: the answer, certainty score (1-5), and a justification in English. The calculations happening beyond that are deterministic.

3.2 Complexity Ceiling

3.2.1 Definition

The concept of a state of complexity beyond which LLMs fail to produce reliable outcomes is already present in the recent scientific literature. Shojaee et al. (2025) identify a complexity cliff in deterministic puzzle environments: Tower of Hanoi, River Crossing, and similar tasks. They found out that

when complexity is gradually increased, the performance of reasoning models improves steadily up until a point at which it "falls off a cliff". Susnjak et al. (2024) define a complexity threshold in a more general sense, in a simulated setting. They also come to the conclusion that there is a point beyond which excess complexity becomes a systemic barrier to reliable model performance. Both studies are concerned with the point of breaking, and both measure proximity to that point through accuracy against a fixed correct answer, either real or simulated. Our case differs on both of these points. First, we are interested in finding the spot right before excess complexity breaks performance. Second, the output is a set of 11 ordinal scores across politically charged text, where correctness cannot be verified by comparison to a fixed reference. In our case, we need to find such a state of balance between internal and external complexity that:

$$c^* = max\{c \in C | S(c) \geq \tau\}$$

Where:

c^* – the complexity ceiling (the optimal configuration)
c – a configuration in the complexity space C, defined by the vector of internal parameters (hyperparameters) and external parameters (context load)
C – the full complexity space being searched
$S(c)$ – the stability function evaluated at configuration c
τ – the stability threshold

In practice, C is a discretized and partially explored subset of the full configuration space, and c^* should be interpreted as a local optimum within this bounded search.

3.2.2 Complexity Types

External Complexity External complexity is defined by context. It is important to note that, although the texts being analysed are a very important part of the context themselves, we did not manipulate them at all. Regardless of the complexity configurations, we always analyse one full text per LLM call. In terms of complexity manipulation, three context engineering strategies were applied.

Batch Prompting Batch prompting (Lin et al., 2024) makes LLM prompting more efficient by combining instructions in fewer, bigger prompts. The authors acknowledge that from a certain point on, the more tasks (complexity) are added to a prompt, the harder it gets for the model to perform well.

The 62 questions were clustered together in two meaningful ways: based on methodology (IWH, FABLE), and based on methodological dimension.

- **Methodology-based:** every text was analysed with a total of 2 prompts, with each of them containing 32 (IWH) and 30 (FABLE) questions.
- **Dimension-focused:** every text was analysed by a total of 11 prompts, with each of them containing between 4 and 15 questions. The Believability dimension of FABLE contains 15 questions, as it is the only dimension where questions carry negative weights.

In-context Learning We employed an in-context learning approach (Wang et al., 2024), in which the input prompt is conditionally parameterized with dimension- or methodology-specific orienting guidelines prior to each inference batch. The guidelines were taken from general guideline documents for human critical discourse analysis (CDA) coders. This was planned as a measure to ease complexity and provide guidance. All configurations in this study are zero-shot — no labelled examples are provided to the model. The orienting guidelines therefore serve as the primary scaffolding for task framing in the absence of demonstrations.

System-level Instructional Conditioning We also put the same guidelines in the system prompt, again conditionally invoked, based on the specific methodology or dimension instructions the input prompt contains (Cheng and Mastropaolo, 2026).

Internal Complexity

Model We worked entirely with Gemini 3 Flash. At the time of selection, it was the most affordable frontier reasoning model reaching above 90% on both Multilingual Q&A (MMMLU) and Commonsense reasoning across 100 Languages and Cultures (Global PIQA) scores (Google DeepMind, 2025). While the approach is model-agnostic in design, its empirical behavior may vary across reasoning LLM architectures and requires validation beyond the single model used in this study.

Thinking level Gemini 3 has four thinking level options - "high", "medium", "low" and "minimal". These levels form a discrete, ordered internal complexity axis, which is just the comfortably

controllable parameter our search for a complexity ceiling requires. The lower the thinking level, the lower the internal complexity load.

Temperature Temperature governs the entropy of the model's output distribution: higher values flatten the probability distribution over possible tokens, increasing lexical and inferential diversity at the cost of determinism. Inversely, lower values concentrate probability mass, producing more focused but less exploratory outputs.

Together, the internal and external parameters defined above constitute the bounded configuration space C for this study. The leading configuration identified as described in 3.3 is therefore a **local optimum within this space**. As it is constrained by the available hyperparameter structure of the selected model and the computational budget, it should not be interpreted as a statement about the global capability limits of the model or of reasoning LLMs in general.

3.3 Descent to Feasibility

3.3.1 Definition

In contrast with the complexity cliff method (Shojaee et al., 2025), where complexity is gradually increased until reaching a point-break, we search for the ceiling in the opposite way. We set the complexity parameters high and gradually decrease them until the stability criteria are met:

$$c_{t+1} = c_t - \Delta c \text{ until } S(c_{t+1}) \geq \tau$$

Where Δc is the reduction step across the complexity dimensions. The algorithm terminates at $c^* = c_{t+1}$, the first point where the feasibility condition is satisfied.

The non-monotonic nature of LLM reasoning under increasing complexity loads (Shojaee et al., 2025) renders bottom-up search strategies suboptimal for identifying the global complexity ceiling. In a non-monotonic stability landscape, incremental ascent risks terminating at the first locally stable configuration encountered - a low-complexity point that satisfies the stability criterion but falls short of a genuinely demanding pipeline. By contrast, descending from a reasonably high point of complexity and stopping at the first stable configuration increases the likelihood that the identified configuration reflects a higher-complexity reasoning regime compared to bottom-up search strategies.

3.3.2 Process

After we define the highest starting point in terms of complexity within the constrained search space, the gradual descent of complexity parameters starts in the following order:

Thinking level The first variable to be reduced is the model's thinking budget — the parameters governing the depth and allocation of abstract reasoning prior to output generation. This ordering is motivated by the interaction between extended reasoning and output instability: high reasoning token allocations, when combined with complex instruction sets and elevated temperature, have been shown to amplify output variance and reduce cross-run modal consistency. Accordingly, thinking parameters are treated as the highest-leverage source of instability and are addressed first, with all other parameters held constant.

Context If the stability criterion remains unmet following reduction of thinking parameters, the process moves on to the context. We don't go with the smaller batch prompts immediately - before that we try to ease complexity by adding general guidelines first to the input prompt, then to the system prompt. Only after this doesn't work out, dimension-specific prompts are employed. This decomposition significantly reduces the simultaneous inferential load imposed on the model without altering the analytical scope of the task. With the smaller prompts, the cycle repeats - starting from the highest thinking-level point, and going down. Bigger batch prompts are more time and energy efficient - they need fewer calls per analysis, therefore our search strategy starts with lowering thinking parameters rather than instruction load.

Temperature Temperature is treated as a fixed anchor throughout stages Thinking level and Context and reduced last. This design serves two purposes. First, it isolates the effect of each descending step: holding temperature constant ensures that stability gains are attributable to reductions in thinking level or context complexity specifically, rather than to a more constrained output distribution (Atil et al., 2025). Second, elevated temperature keeps both complexity dimensions genuinely active simultaneously — reducing it prematurely would suppress the reasoning process rather than calibrate it, producing a spurious stability that reflects reduced expressiveness rather than genuine pipeline robustness. Temperature reduction is therefore reserved

for cases where the full descent at higher entropy fails to yield a shortlisted configuration.

3.4 Stability Benchmark

Here we start from the assumption of Atil et al. (2025), which states that stability depends on the specific combination of task type, output format, and prompt structure. In our case, this means that instability in Bulgarian high-subtext scoring is not predictable from general benchmarks. We needed a systematic way to measure it directly.

3.4.1 Definition - 80/80 rule

At least 80% of all texts sampled for the optimal pipeline search should produce the same final score at least 8 times after 10 independent runs across all 11 dimensions, provided that modal consistency for all texts across all dimensions averages above 80%. For every dimension $d \in D$, we calculate the stability score as follows. Given:

N – total number of texts in the pilot sample
$R = 10$ – number of independent runs per text
$s_{i,d,r} \in \{0, 1, 2\}$ – the score assigned to text i on dimension d in run r
$\hat{s}_{i,d}$ – the modal score of text i on dimension d across all R runs:

$$\hat{s}_{i,d} = \text{argmax}_{v \in \{0,1,2\}} \sum_{r=1}^{R} \mathbf{1}\{s_{i,d,r} = v\}$$

A text i is modally consistent on dimension d if its modal score appears in at least 8 out of 10 runs:

$$M_{i,d} = 1\Big\{ \sum_{r=1}^{R} 1\{s_{i,d,r} = \hat{s}_{i,d}\} \geq 0.8R \Big\}$$

The dimension-level stability score is then:

$$S_d = \frac{1}{N} \sum_{i=1}^{N} M_{i,d}$$

3.4.2 Rationale

Why 10 Runs According to Atil et al. (2025), and Wang and Wang (2025), 5 runs are enough to get most information about LLM output stability, with the latter study showing that every subsequent run beyond that has considerably flattening returns.

Why Modal Consistency Although consistency and accuracy are clearly two distinct terms, Wang et al. (2023) demonstrated empirically that the most common outcome from different chain-of-thought reasoning pipelines has much higher probability to be the correct one, linking modal consistency across different complex thinking processes to higher accuracy. In our case, higher temperature and thinking levels ensure that every run produces new, different reasoning patterns. However, as ground truth is not directly observable, we treat stability under a heavy complexity load strictly as a **proxy** rather than a measure of correctness.

Why 80/80 80% is a standard reliability benchmark across content analysis and (Neuendorf, 2002; McHugh, 2012; Lombard et al., 2002). In light of the intrinsic ambiguity of politically charged Facebook posts in Bulgarian, it's unrealistic to expect that all texts can be clearly defined according to the standards of all 11 dimensions. Therefore, we decided to loosen-up the stability criteria in that regard to 80% of the texts being required to pass this threshold. To further address concerns about statistical precision at small sample size, we add the requirement of at least 80% of average modal consistency for all texts, across all dimensions. When the central tendency of the distribution is itself above the criterion value, the conclusion is robust to sample size variation - large effects are detectable even at small N (Cohen, 1988). Furthermore, our 80/80 rule has an analogical precedent in a famous approach for the stability assessment of another area where higher-order reasoning is concerned - the educational process. Block and Anderson (1975) have assumed that a learning mastery process can be considered stable if 80% of the students pass 80% of the criteria on every single subject.

4 Results

4.1 Search Space and Algorithm

We have defined a search space for this study, based on the complexity parameters described above, which can be seen in Table 1.

Due to resource restrictions, configurations under temperature of 0.8 were not fully tested. A total of 14 configurations were tested following Appendix B: Algorithm 1, with 2 other configurations tested outside of this scope for referential purposes. The results for all tested configurations are listed in Appendix A: Table 2.

Eventually, three of the configurations with temperature set at 1 came very close to meeting the 80/80 criterion (Figure 1, Appendix C: Figure 2).

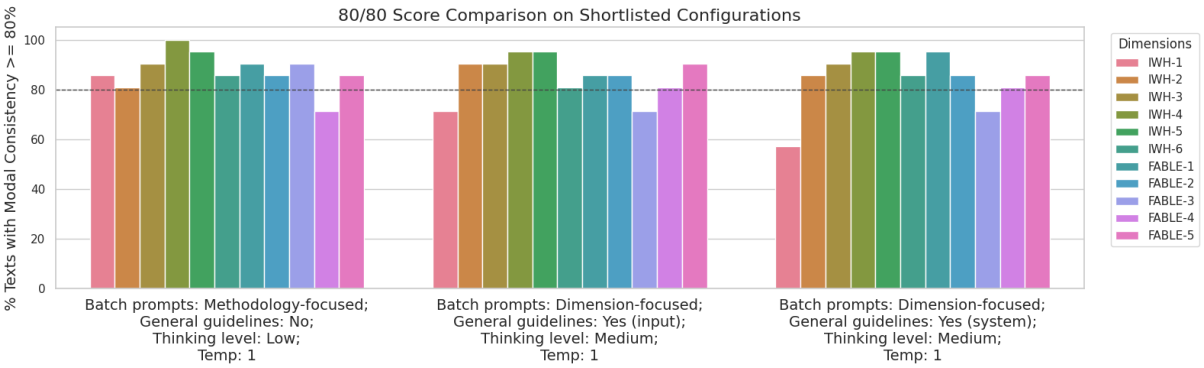

Figure 1: Comparison between the shortlisted configurations, returned by Algorithm 1.

Complexity Parameter	Values
Batch Prompting	Methodology-focused (big), Dimension-focused (small)
General Guidelines	No, Yes (input prompt), Yes (system prompt)
Thinking Level	Medium, Low
Temperature	1.0, 0.8

Table 1: Search space for the tested complexity configurations.

4.2 Problematic Dimensions

The shortlist threshold of $k = 9$ reflects two empirically grounded observations: first, IWH-1 exhibited instability across the majority of tested configurations, suggesting genuine interpretive ambiguity in this dimension; second, FABLE-3's instability under dimension-focused prompts is attributable to its disproportionate question load (15 questions vs. 4-6 for all other dimensions), making it a structural rather than semantic source of variance. Requiring 11/11 stability would therefore conflate pipeline performance with dimension-specific factors.

4.3 Shortlisted Configurations

Three configurations were shortlisted - one with methodology-focused batch prompts, and another two with dimension-focused ones.

4.3.1 Methodology-focused Batch Prompts

The complexity configuration with highest stability had methodology-specific batches, "low" thinking level and temperature of 1, when the batches were put into the system prompt (Figure 1). Exactly the same configuration produced clearly less stable results, when the batch prompts were put into the input prompt part (Appendix D: Figure 3).

4.3.2 Dimension-focused Batch Prompts

Two nearly identical configurations came close to fully meeting the 80/80 criteria as well (Figure 1), differing only in whether the guidelines were placed in the input prompt or the system prompt. Both had 9 out of 11 dimensions pass our stability threshold, suggesting that placement of general guidelines is not a decisive stability factor at this complexity level. Here, the annotation guidelines came on top of a pipeline that failed on 3 out of 11 dimensions and directly improved stability on IWH-2 in both cases.

More dimension-specific configurations satisfied the 80/80 standard with the same success rate (9 out of 11) but they were not shortlisted due to their lower levels of thinking complexity.

5 Discussion

Balance Between Internal and External Complexity The results point to two structurally distinct near-ceiling configurations, each reflecting a different balance between internal and external complexity. The first concentrates complexity in the context through methodology-focused batch prompts, while keeping the thinking level low. The second spreads context load across smaller, dimension-focused prompts and compensates with a higher reasoning budget at medium thinking level.

These configurations do not merely differ in parameter values; they appear to excel at different analytical tasks. The methodology-focused, low-thinking configuration achieves stable results not only on dimensions with relatively direct textual evidence, but also on IWH-1 and FABLE-3, which proved the hardest to stabilize overall. A plausible explanation is that a broader contextual frame combined with constrained inferential depth reduces the model's tendency to over-interpret inherently

ambiguous dimensions.

The dimension-focused configurations, by contrast, appear better suited to dimensions whose indicators are not explicitly stated in the text but require higher-order inference. FABLE-3 (Believability) remains a consistent instability point for this configuration type, but we attribute this primarily to its disproportionate question load: 15 questions compared to 4 to 6 for every other dimension. This interpretation is supported by the strong stability performance of all other dimension-specific prompts under the same configuration (except for IWH-1), suggesting that medium-thinking pipelines can achieve stable outputs under constrained prompt complexity.

The Role of General Guidelines When we put general-level guidelines, the prompt-heavy configurations' stability declined consistently. The same measure had, also sustainably, the opposite effect when applied to the smaller dimension-specific prompt configurations. We can hypothesise that a possible explanation here could be that when complexity is predominant in the context, adding steering context adds even more complexity, thus causing even more instability. This observation needs further research for a precise explanation, but it demonstrates the subtle balance between external and internal complexity.

6 Conclusion

While the approach is not inherently language-specific, its applicability to other low-resource environments remains an open empirical question. Politically charged and culturally specific narratives will always require high reasoning capacity and stability. Not a trade-off, or a compromise - both of them. This work demonstrates that guiding a reasoning LLM with strong multilingual capacity from a high-complexity state toward a stable configuration is a viable strategy for achieving both analytical depth and scoring consistency.

However, we should also acknowledge this study's **limitations**. This study has used only one reasoning model, a constrained search space, and a single political event in one low-resource language. Broader empirical validation across LLMs, larger search spaces, political contexts, and languages is needed before the approach can be considered fully operational. Importantly, this study evaluates stability independently of ground-truth correctness.

Future work includes optimizing the search for complexity ceiling with a larger search space, adapting the methodology to non-political use cases and other low-resource languages. We also plan to pair the 80/80 criterion with human-annotated validity checks and other sources or approximations of ground truth. Last but not least, we are already working on applying this methodology to multimodal media content.

References

Salah Feras Alali, Mohammad Nashat Maasfeh, Mucahid Kutlu, and Saban Kardas. 2026. Measuring political stance and consistency in large language models.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" LLM settings.

James H. Block and Lorin W. Anderson. 1975. *Mastery Learning in Classroom Instruction*. Macmillan, New York.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Zaiyu Cheng and Antonio Mastropaolo. 2026. An empirical study on the effects of system prompts in instruction-tuned models for code generation.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates.

Roberts Darģis, Guntis Bārzdiņš, Inguna Skadiņa, Normunds Grūzītis, and Baiba Saulīte. 2024. Evaluating open-source LLMs in low-resource languages: Insights from Latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293, Miami, USA. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 flash: Model card and evaluation. Accessed: 2026-04-15.

Antonio Gramsci. 1971. *Selections from the Prison Notebooks of Antonio Gramsci*. International Publishers, New York.

Ömer Alperen Gürses and İsmail Ceylan. 2026. Consistency over accuracy: run-to-run stability of contemporary large language models on turkish curriculum-aligned theoretical anatomy multiple-choice questions. *BMC Medical Education*, 26:285.

Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. Large language models are better logical fallacy reasoners with counterargument, explanation, and goal-aware prompt formulation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6933–6952, Albuquerque, New Mexico. Association for Computational Linguistics.

Keith Peter Kiely. 2025. Wearing gramsci's glasses: transition, conspiracy and disinformation in bulgaria through the lens of hegemony. *Journal of Political Ideologies*, 0(0):1–20.

Jianzhe Lin, Maurice Diesendruck, Liang Du, and Robin Abraham. 2024. Batchprompt: Accomplish more with less.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604.

Pinja M. Marin, Marjaana Lindeman, and Annika M. Svedholm-Häkkinen. 2024. Susceptibility to poor arguments: The interplay of cognitive sophistication and attitudes. *Memory & Cognition*, 52(7):1579–1596.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

Anushka Manchanda Nair. 2025. Multi-stage LLM reasoning for automated detection and classification of high-impact misinformation. Master's thesis, Massachusetts Institute of Technology.

Kimberly A. Neuendorf. 2002. *The Content Analysis Guidebook*. Sage, Thousand Oaks, CA.

Oana Olariu. 2022. Critical thinking as dynamic shield against media deception: Exploring connections between the analytical mind and detecting disinformation techniques and logical fallacies in journalistic production. *Logos Universality Mentality Education Novelty: Social Sciences*, 11(1).

Ketan Patel, Christian Hansmeyer, Nandan Desai, and Aditya Ajit. 2025. American hegemony at a critical juncture, lessons from history's great powers. *Frontiers in Political Science*, Volume 7 - 2025.

Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X. Zhang. 2024. Misinformation as a harm: Structured approaches for fact-checking prioritization. *arXiv preprint arXiv:2312.11678*.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Teo Susnjak, Timothy R. McIntosh, Andre L. C. Barczak, Napoleon H. Reyes, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Over the edge of chaos? excess complexity as a roadblock to artificial general intelligence.

Julian Junyan Wang and Victor Xiaoqi Wang. 2025. Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Appendix A Table 2: 80/80 scores for each configuration tested

#	Batch type	General Guidelines	Thinking Level	Temperature	Dimensions Passing
1	Methodology-Focused	No	Medium	1.0	7/11
2	Methodology-Focused	Yes (Input)	Medium	1.0	5/11
3	Methodology-Focused	Yes (System)	Medium	1.0	6/11
4	Methodology-Focused	Yes (System)	Medium	0.8	6/11
5	Methodology-Focused	No	Low	1.0	10/11
6	Methodology-Focused	Yes (Input)	Low	1.0	7/11
7	Methodology-Focused	Yes (System)	Low	1.0	7/11
8	Methodology-Focused	Yes (System)	Low	0.8	8/11
9	Dimension-Focused	No	Medium	1.0	8/11
10	Dimension-Focused	Yes (Input)	Medium	1.0	9/11
11	Dimension-Focused	Yes (System)	Medium	1.0	9/11
12	Dimension-Focused	No	Low	1.0	9/11
13	Dimension-Focused	Yes (Input)	Low	1.0	9/11
14	Dimension-Focused	Yes (System)	Low	1.0	7/11

Table 2: 80/80 results for each configuration tested

Appendix B Algorithm 1

Algorithm 1 Descent to Feasibility

Input: $\tau = 0.80$, $R = 10$, acceptance threshold $k = 9$, D: set of dimensions ($|D| = 11$), $S_d(c, R)$: stability score of dimension d under configuration c across R independent runs

Output: *shortlist*: shortlist (all c with passed $\geq k$); c^* if full ceiling found

shortlist $\leftarrow \varnothing$
$c^* \leftarrow \varnothing$
for $temp \in \{1.0, 0.8\}$ **do**
  for $batch \in \{\text{Methodology-Focused}, \text{Dimension-Focused}\}$ **do**
    for $guidelines \in \{Input\_Prompt, System\_Prompt\}$ **do**
      for $thinking \in \{Medium, Low\}$ **do**
        $c \leftarrow (batch, guidelines, thinking, temp)$
        $passed\_count \leftarrow |\{d \in D \mid S_d(c, R) \geq \tau\}|$
        if $passed\_count = |D|$ **then**
          $shortlist \leftarrow shortlist \cup c$
          $c^* \leftarrow c$
          return $shortlist, c^*$
        else if $passed\_count \geq k$ **then**
          $shortlist \leftarrow shortlist \cup c$
return $shortlist, c^*$

Appendix C Figure 2: Average Modal Consistency

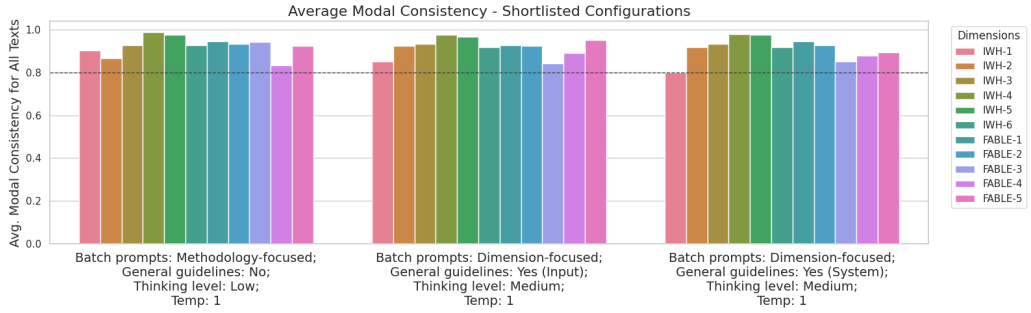

Figure 2: Average modal consistency comparison between the three shortlisted configurations.

Appendix D Figure 3: Leading Config with Batch Prompts in System Prompt vs Batch Prompts in User Prompt

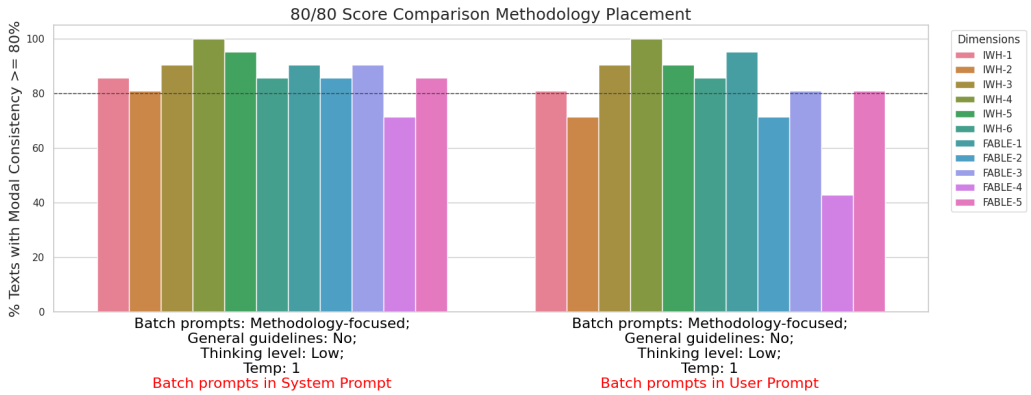

Figure 3: Comparison between placing the methodology in the system prompt, as opposed to the user prompt. The other configurations remain the same.

Appendix E Coding Questionnaire

Below is the complete list of questions used in the evaluation phase of the study:

IWH-1: Fragmented Consensus (5 items):
Q1. Does the text highlight polls or anecdotes that Bulgarians are deeply divided on the euro? (Yes/No/Unclear)
Q2. Does it contrast specific demographic groups (urban vs rural, educational levels, wealthy vs poor, young vs old, elite vs people)? (Yes/No/Unclear)
Q3. Does it cite recent electoral volatility linked to the currency issue? (Yes/No/Unclear)
Q4. Are coalition rifts over the euro portrayed as irreconcilable? (Yes/No/Unclear)
Q5. Is the societal split itself framed as a central crisis that must be addressed? (Yes/No/Unclear)

IWH-2: Counter-Hegemonic Forces (5 items):
Q6. Are organized opposition actors (petition committees, civic groups) spotlighted? (Yes/No/Unclear)
Q7: Does the content call for or depict a specific, organized action (e.g., a named protest, a petition with a link)? (Yes/No/Unclear)
Q8: Does the content use strong rhetorical calls for general resistance or civil disobedience (e.g., "resistance is a duty")?
(Yes/No/Unclear) Q9. Are alternative media sources or Telegram channels presented as information hubs? (Yes/No/Unclear)
Q10. Is the counter-elite narrative presented as gaining momentum or influence? (Yes/No/Unclear)

IWH-3: Crisis of Authority (4 items):
Q11. Does it reference repeated elections, cabinet collapses, or leadership resignations? (Yes/No/Unclear)
Q12. Are court battles (e.g., Constitutional Court decisions) cited as blocking policy implementation or a major democratic process (such as a referendum or popular initiative)? (Yes/No/Unclear)
Q13. Is low public trust in parliament, BNB, or government ministries emphasized? (Yes/No/Unclear)
Q14. Does it characterize current leaders as "illegitimate" or "without mandate"? (Yes/No/Unclear)

IWH-4: Reliance on Coercion (3 items):
Q15. Are state-aligned outlets accused of censoring opposition voices or presidential speeches? (Yes/No/Unclear)
Q16. Is there mention of legal threats, fines, or systematic smear campaigns against dissenters? (Yes/No/Unclear)
Q17. Does the content claim that organized political opposition (e.g., parties, formal movements) are being systematically silenced or censored? (Yes/No/Unclear)

IWH-5: Ideological Contradictions (4 items):
Q18. Does it accuse anti-euro elites of maintaining euro bank accounts abroad while opposing adoption? (Yes/No/Unclear)
Q19. Does it highlight MPs pricing real estate in euros while rejecting public euro debate? (Yes/No/Unclear)
Q20. Are governing parties portrayed as pro-EU publicly but corrupt or hypocritical privately? (Yes/No/Unclear)
Q21. Are there examples of sudden policy U-turns ("a referendum was unnecessary, now it is acceptable")? (Yes/No/Unclear)

IWH-6: Moral/Intellectual Void (5 items):
Q22. Does it deride economists or academics as "Soros puppets" or "on the payroll"? (Yes/No/Unclear)
Q23. Is expert analysis replaced by conspiracy language or populist slogans? (Yes/No/Unclear)
Q24. Are complex fiscal issues reduced to emotive metaphors (e.g., euro slavery)? (Yes/No/Unclear)
Q25. Does it lament the absence of authentic leaders" with principles? (Yes/No/Unclear)
Q26. Are appeals made to raw emotion over evidence ("voice of people, voice of God")? (Yes/No/Unclear)

FABLE-1: Fragmentation (6 items):
Q27. Does the post use a rhetorical frame that divides society into two opposing, irreconcilable groups (e.g., "the people vs. the elite," "patriots vs. traitors")? (Yes/No/Unclear)
Q28. Does it portray Bulgarian state institutions as fundamentally untrustworthy? (Yes/No/Unclear)
Q29. Does it attack mainstream Bulgarian media credibility systematically? (Yes/No/Unclear)
Q30. Does it question the honesty or patriotism of "ordinary Bulgarians" who disagree? (Yes/No/Unclear)
Q31: Does the post claim that the outcome of an election/referendum/constitutional court decision was (or will be) fraudulent or rigged? (Yes/No/Unclear)
Q32: Does the post claim that the process for initiating or conducting an election/referendum was illegally blocked or sabotaged? (Yes/No/Unclear)

FABLE-2: Actionability (6 items):
Q33. Is there an explicit call to sign petitions, share content, protest, or block legislation? (Yes/No/Unclear)
Q34. Are specific dates, places for offline action, or livestream details provided? (Yes/No/Unclear)
Q35. Does it urge immediate action with urgency language ("now", "before it's too late")? (Yes/No/Unclear)
Q36. Does it single out specific individuals or offices for pressure or harassment? (Yes/No/Unclear)

Q37. Does it link euro grievances to vulnerable groups (children, pensioners) to mobilize protection instincts? (Yes/No/Unclear)
Q38. Is the message part of a coordinated campaign or series (organized hashtags)? (Yes/No/Unclear)

FABLE-3: Believability (15 items):
High-Risk Believability Factors (Increase Score):
Q39. Does it rely primarily on personal anecdotes or testimonials as evidence? (Yes/No/Unclear)
Q40. Does it use "everyone knows" or appeal to common sense? Yes/No/Unclear)
Q41. Does it present slippery slope arguments ("euro will lead to complete sovereignty loss")? (Yes/No/Unclear)
Q42. Does it exploit knowledge gaps ("experts won't tell you this")? (Yes/No/Unclear)
Q43. Does it make causal claims without proof ("euro caused Greece crisis")? (Yes/No/Unclear)
Q44. Is the primary format narrative/storytelling rather than factual? (Yes/No/Unclear)
Q45. Does it target audiences with lower formal education/media literacy? (Yes/No/Unclear)
Q46. Is sharing motivated by identity expression or persuasion? (Yes/No/Unclear)

Neutral Believability Factors:
Q47. Does it mix factual information with selective evidence? (Yes/No/Unclear)
Q48. Is it published by outlet with unclear editorial standards? (Yes/No/Unclear)

Low-Risk Believability Factors (Decrease Score):
Q49. Does it rely heavily on ad hominem attacks? (Yes/No/Unclear)
Q50. Does it contain obvious logical contradictions? (Yes/No/Unclear)
Q51.Does it use extreme polarization that may trigger skepticism? (Yes/No/Unclear)
Q52. Does it lack any supporting evidence whatsoever? (Yes/No/Unclear)
Q53. Is the tone so extreme it appears fabricated? (Yes/No/Unclear)

FABLE-4: Likelihood of Spread (5 items):
Q54. Is the author an influencer, major party page, or national TV/media site with significant reach? (Yes/No/Unclear) Note: Code based on the immediate publisher of the content being analyzed. If a personal blog quotes a national TV station, the author is the blogger, not the station.
Q55. Is the content published on high-traffic, public platforms (Facebook, TikTok, YouTube)? (Yes/No/Unclear)
Q56. Are viral elements present (hashtags, mentions, paid promotion, or boosting evidence)? (Yes/No/Unclear)
Q57. Does it contain arresting visuals, caps-lock headlines, emojis, or eye-catching formatting? (Yes/No/Unclear)
Q58. Is it tied to active news cycles (e.g., Radev's speeches, parliamentary debates, EU announcements)? (Yes/No/Unclear)

FABLE-5: Exploitativeness (4 items):
Q59. Does it warn of imminent threats (poverty, sovereignty loss, economic "robbery") to generate fear? (Yes/No/Unclear)
Q60. Does it emphasize powerlessness with specific examples ("160 MPs rob the people")? (Yes/No/Unclear)
Q61. Does it specifically target groups with fewer resources or higher vulnerability (rural populations, elderly)? (Yes/No/Unclear)
Q62. Is the content exclusively in Bulgarian, targeting monolingual audiences with limited cross-referencing ability? (Yes/No/Unclear)