**The reason for AI-prompting standard**

To investigate the effects of scapegoating rhetoric, this study operationalized textual stimuli into two high-contrast categories—scapegoating misinformation and true information—based on a calculated "Scapegoating Score." This score was derived by segmenting each text into its constituent clauses and determining the percentage that attributed complex problems to a single entity (e.g., an individual, party, or group). We established a dichotomous framework wherein "scapegoating misinformation" was defined as texts with a Scapegoating Score of 70 or greater, while "true information" was defined as texts with a score of exactly 0. This 70/0 threshold was deliberately chosen to maximize the experiment's internal validity and statistical power. The high threshold of ≥70 ensures that the treatment condition is a potent and unambiguous manipulation, composed of texts where the scapegoating narrative is pervasive rather than incidental, thus strengthening the construct validity. Conversely, the stringent criterion of a zero score for the control condition creates a methodologically pure baseline, guaranteeing the complete absence of the rhetorical device under investigation. This allows for a clear causal inference by comparing the manifest presence of scapegoating with its absolute absence. Consequently, all texts with scores from 1 to 69 were excluded. This large exclusionary zone serves to eliminate ambiguous or mixed-message stimuli that could introduce statistical noise and increase within-group variance. By creating maximally distinct stimulus groups, this framework enhances the signal-to-noise ratio, providing a rigorous and powerful test of our research hypothesis.

**Prompt for AI of “scapegoating” misinformation discrimination**

<<<BEGIN PROMPT>>> You are a text analyst tasked with evaluating the "Scapegoat Proportion" of an input text according to the following operational rules. Your output should be a final score from 0–100 and a detailed analysis log.

I. Core Definition

In this task, "Scapegoating" refers to the oversimplification of issues with potentially multiple, complex causes into a narrative where a single person, a single political party, a single group, or a single external entity is presented as the sole cause, while ignoring other potential factors or one's own responsibility. This is often manifested through negative labeling, "culprit-finding" language, and the attribution of multiple different negative outcomes to the same single entity.

[Key Reminder] If the text also mentions multiple causes, acknowledges partial self-responsibility, or lists multiple variables in an analytical tone, the "Scapegoat intensity" should be reduced.

II. Judgment Indicators (Clause-by-Clause Check)

For each "clause" (see splitting rules in the next section), check if it meets any of the following five indicators. Each hit counts as 1 point (for a maximum of 5 points per clause):

(1) Single Culprit Language: Directly accuses or implies that a specific entity is the main culprit, the root cause, the source of the trouble, or the problem's core (including phrases semantically equivalent to an accusation, e.g., "It's all X's fault," "X is the real problem"). (2) Complex-to-Single-Cause Simplification: Reduces structural or complex issues (e.g., economy, energy, diplomacy, social conflicts) to a single entity's fault; or uses exclusionary phrasing like "the only reason is..." or "it's all because of..." to attribute cause. (3) Negative Emotion / Stigmatizing Labels: Uses emotionally negative labels to reinforce blame (e.g., idiot, shameless, traitor, selling out the country, killing us all, garbage...). The function of these labels is to stigmatize the target to facilitate blame attribution. (4) Ignoring Other Factors and Self-Blame: When discussing a problem, the text only attacks an external entity without mentioning its own role, systemic factors, or other potential sources; or the phrasing leads the reader to believe there is no second cause. (5) Linking Multiple Consequences to the Same Entity: Lists multiple distinct negative outcomes (e.g., tariffs, energy, public welfare, diplomacy, security...) and claims they are all caused by the same single entity (i.e., "all-encompassing" blame attribution).

Note: A single clause can hit multiple indicators. Clauses that are merely stating background facts, quoting multiple viewpoints, or are neutral in tone should not be scored.

III. Clause Splitting Rule

First, split the input text into a "list of clauses" to facilitate itemized annotation:

1. Use periods (.), exclamation marks (!), question marks (?), and paragraph breaks as the first layer of separators.
2. If a single sentence contains multiple attributions of blame or multiple claims, further split it at clear semantic breakpoints like commas (,) or semicolons (;).
3. Each resulting clause should be able to stand alone as "one claim" or "one entity + one evaluation/causal link."
4. Control: If splitting a clause too finely makes its meaning unclear, it is better to keep the clause longer. However, when encountering a clear serial list of blames, it is mandatory to split them.

IV. Scoring Steps

For each clause:

* Record the numbers of the indicators that were hit (1–5).
* Clause Score = (Number of hit indicators) / 5 (range: 0–1).
* Clause Binary Flag = 1 if ≥1 indicator is hit, otherwise 0.

For the entire text:

* N = Total number of clauses.
* Binary\_Proportion = (Σ Clause Binary Flags) / N.
* Weighted\_Proportion = (Σ Clause Scores) / N.
* Final\_Scapegoat\_Score\_0to100 = round(Weighted\_Proportion \* 100).
* Also, report the Binary\_Proportion for comparison (can be presented as a percentage by multiplying by 100).

V. Output Format

Produce the output according to the following structure (you must follow the order and headings; use plain text, not tables; line breaks are acceptable):

[Scapegoat Score Result] Total Clauses: N = <number> Hit Clauses: <number> (Binary\_Proportion = <decimal, 0-1>; Percentage = <rounded percentage>%) Weighted Proportion = <decimal, 0-1> Final\_Scapegoat\_Score\_0to100 = <integer, 0-100>

[Clause-by-Clause Analysis]

1. <Clause text> | Hit Indicators: <List 1,3,5... or None> | Clause Score=<0-1>
2. <Clause text> | ... (List all clauses sequentially, maintaining their original order.)

[TEXT\_TO\_EVALUATE] <<<END PROMPT>>>

**The reason for human coding rubric**

This study employed a four-point ordinal scale (0-3) for the human coding of scapegoating rhetoric intensity in social media posts. The design of this protocol was intended to establish a measurement tool that possesses high reliability, internal validity, and ecological validity. Each component of this protocol, from the unit of analysis to the specific definitions of the scale, is grounded in rigorous methodological considerations to ensure the robustness and replicability of the findings.

First, the decision to use the entire post as the unit of analysis is based on the principle of ecological validity. The rhetorical effect of scapegoating is often achieved through contextual framing, emotional escalation, and the cumulative impact of multiple arguments. Dissecting a post into smaller units might sever its semantic context and lead to a misjudgment of its true persuasive intent. A holistic assessment of each post most closely simulates how users naturally receive and interpret information, thereby accurately capturing the overall intensity and primary thrust of its narrative.

Furthermore, the adoption of a four-point ordinal scale, rather than a direct binary (Yes/No) classification, was intended to enhance measurement precision and inter-rater reliability. A simple binary choice forces coders to make an absolute judgment on ambiguous texts of varying intensity, which can amplify minor perceptual differences between raters and consequently lower reliability. The four-point scale provides a more nuanced framework. The score of '0' represents a qualitative 'true zero point,' clearly defining a baseline for the absence of the phenomenon. This ensures that irrelevant content, such as objective factual statements, is consistently excluded, avoiding the measurement error of conflating 'low intensity' with 'non-existence'. The tiers from 1 to 3 reflect three distinct qualitative levels of scapegoating rhetoric: from an implicit suggestion, to an explicit accusation, and finally to an extreme escalation. These levels are not merely arbitrary quantitative steps but represent clearly distinguishable qualitative shifts, providing coders with stable and operational anchors for their judgments.

The definition of the '3' score (extreme scapegoating) is grounded in established theories of rhetoric and political communication. The designation of 'stigmatizing labels,' 'over-generalization of responsibility,' and 'absolutist language' as its determining 'aggravating factors' is based on their recognition in academic literature as high-intensity propaganda techniques used to polarize debate, simplify attribution, and delegitimize opponents. This definition ensures that the highest score is awarded not on the basis of subjective perceptions of intensity, but on the presence of objective, identifiable rhetorical strategies, thereby significantly increasing the objectivity of the coding standard.

Finally, the conversion of the 0-3 ordinal ratings into a binary 'Yes/No' classification serves the study's statistical and experimental design requirements. While the ordinal scale is crucial during the measurement phase, inter-rater reliability statistics (e.g., Fleiss' Kappa) and the subsequent experimental operationalization of posts into 'treatment' (containing scapegoating) and 'control' (not containing scapegoating) groups necessitate a clear categorical division. The logic of converting a score of '0' to 'No' and scores of 1-3 to 'Yes' is that once a text crosses the threshold from absence to presence—even if merely implicit—it qualitatively falls within the category of scapegoating rhetoric. The detailed four-point scale thus functions as a precision calibration tool to ensure that this final binary classification is highly reliable and consistent.

In summary, this coding protocol—through its use of holistic assessment, a true zero point, and theory-based tier definitions—systematically addresses the ambiguity and subjectivity inherent in measuring complex social science constructs. It thereby provides a logically coherent and operationally robust quantitative foundation for the study's core variable.

**Rubric for human of “scapegoating” misinformation discrimination**

In this study, "scapegoating rhetoric" is operationalized as a rhetorical strategy that oversimplifies complex, multi-causal public issues by attributing blame to a single actor.

Coders were tasked with reading each designated post and assigning a score from 0 to 3 based on its overall narrative content.

A score of 0 indicates that the post contains no elements of scapegoating. Such posts are characterized by objectivity and neutrality, with content that may include verifiable factual statements, balanced multi-causal analyses, or reasonable critiques of specific, singular events.

A score of 1 is assigned to posts containing implicit scapegoating rhetoric. In these cases, attribution is not made through direct assertion but rather through indirect means such as innuendo, temporal association, or leading language, which guide readers to infer a link of responsibility to a specific target.

A score of 2 denotes posts that feature explicit scapegoating rhetoric. The accusation is direct, clear, and unambiguous, attributing the cause of a complex problem to a single actor using exclusionary logic.

The highest score of 3 is reserved for posts containing extreme scapegoating rhetoric. In addition to meeting the criteria for a score of 2, such posts must also include at least one of the following "aggravating factors": (a) the use of stigmatizing or ad hominem labels; (b) an over-generalization of responsibility, wherein multiple negative outcomes from disparate domains are attributed to the same target; or (c) the use of absolutist language, such as "the root of all evil."

Finally, for the purpose of inter-rater reliability statistics and subsequent analyses, all 0-3 ordinal scores were converted into a binary categorical variable. Posts receiving a score of 0 were categorized as "No" (non-scapegoating rhetoric), while posts receiving a score of 1, 2, or 3 were categorized as "Yes" (scapegoating rhetoric).