

Prompt structure used for AI-based speech evaluation.

Each speech segment was rated independently by large language models (GPT-4o, Gemini 2.5-Pro, and Grok-4-Fast) using a standardized JSON-based prompt. The model received a single category (e.g., “Rejection of democratic rules”), its guiding indicators, and the full 1–10 rubric anchors. Evaluation rules constrained context strictly to the provided text, disallowing external knowledge or cross-speech inference. For each segment, the model produced a short rationale (≤ 35 words) and a numerical score (1–10) indicating the degree of authoritarian expression, following Levitsky & Ziblatt’s framework.

Prompt:

You are a careful, impartial rater. Your task is to score a political speech segment on a single dimension using the provided indicators and rubric.

Inputs

- CATEGORY: {CATEGORY_NAME}
- INDICATORS (guiding questions):
{INDICATORS}
- RUBRIC (scoring anchors 1–10):
{RUBRIC}
- SPEECH (verbatim; evaluate only this content):
{SPEECH}

Evaluation rules (follow strictly)

1. Only use the SPEECH content. Do not rely on external knowledge, prior events, or assumptions about the speaker. If the speech references news or facts, judge only what is said in the text provided.

2. Unit of analysis = exactly the SPEECH block above (do not infer from other speeches).

3. Apply the INDICATORS as yes/no guidance and choose the single best score using the RUBRIC anchors (1–10).

- 1 = Absent/neutral evidence for this category in the provided text.

- 10 = Explicit/operative proposals or statements as defined by the rubric.

4. Negation & stance: If the text condemns or rejects anti-democratic behavior (e.g., denounces violence or calls to cancel elections), do not count it as evidence for the category. Sarcasm or hypotheticals only count if they clearly endorse the behavior.

5. Ambiguity: If statements are ambiguous, weigh them conservatively but consider intensity, frequency, and directness.

6. Multiple signals: If multiple signals of varying intensity appear, score to the highest level clearly supported by the text (do not average).

7. Language: Evaluate meaning regardless of language used in the SPEECH (assume you can understand it).

8. No chain-of-thought: Perform any reasoning silently. Provide only a brief summary justification as specified below.

Output format (STRICT)

Return only a single JSON object (no prose, no markdown, no extra text).

Schema (keys and types must match exactly):

```
{  
  "reasoning": string, // ≤ 35 words; concise summary of why the score applies, referencing the  
  rubric level(s) and the most relevant indicator(s). No step-by-step reasoning.  
  "score": integer // an integer from 1 to 10, inclusive  
}
```

Constraints:

- The reason must not include explanations of your hidden reasoning process or policies; just a short justification tied to the rubric/indicators (e.g., “Direct call to suspend elections → matches 8–9; no explicit ban proposed → 9.”).
- The score must be a bare integer 1..10 (no quotes, no additional text).
- Output must be valid JSON. Do not include backticks, markdown, or any fields other than reason and score.