Codebook for LLM Output Validation Task

1. Introduction and Purpose

This codebook provides guidance for validating the accuracy of automated extractions from parliamentary debate summaries. An LLM has been tasked with extracting four types of information about a specific speaker's contributions: **issues**, **positions**, **arguments**, and **proposals**. Your role is to verify whether these extractions accurately reflect what is communicated in the summary of the debate.

2. Task Overview

For each debate summary, you will:

1. Read the Full debate summary on the left-hand side
2. Evaluate whether the LLM correctly extracted information for **the specified speaker only**
3. Assign a thumbs up (✓) or thumbs down (✗) for each of the four extraction categories (issues, positions, arguments, and proposals)
4. Document your reasoning in the notes field for any decisions that were noteworthy or difficult

Critical Rule: The LLM should ONLY extract information attributed to the **named speaker** identified at the top of the app. Information from other speakers is not relevant.

3. Definitions and Distinctions

3.1 Extracted Issue

What it is: Key problems, concerns, or topics that the speaker raises or identifies as important.

Examples of issues:

- "The inefficiency of current fragmented weapon systems"
- "Europe's dependence on external powers for defence"
- "Rising healthcare costs"
- "Climate change impacts on agriculture"

What it is NOT:

- Solutions or proposals (those belong in Proposals)
- The speaker's opinion on the issue (that belongs in Positions)
- Issues raised by other speakers

3.2 Extracted Position

What it is: The speaker's stance, viewpoint, or opinion on an issue. This expresses *what they believe* or *where they stand on an issue*.

Examples of positions:

- "Europe must become more resilient and reduce dependence on external powers"

- "Supports increased military spending"
- "Opposes the current immigration policy"
- "Believes diplomatic engagement is preferable to military intervention"

What it is NOT:

- Just identifying an issue (that is an Issue)
- Specific solutions (that is a Proposal)
- The reasoning behind their position (that is an Argument)

3.3 Extracted Argument

What it is: The reasoning, evidence, or justification the speaker uses to support their position. This answers *why* they hold their position.

Examples of arguments:

- "Fragmented weapon systems hamper effective defence efforts across EU nations"
- "Historical evidence shows that economic sanctions are more effective than military intervention"
- "Data indicates that prevention programmes reduce long-term costs"

What it is NOT:

- The position itself (that is a Position)
- Proposals for action (that is a Proposal)
- Arguments made by other speakers

3.4 Extracted Proposal

What it is: Concrete suggestions, recommendations, or policy actions the speaker puts forward to address issues. This answers *what should be done*.

Examples of proposals:

- "Establish a European Defence Community to foster greater coordination"
- "Increase defence budget to 2% of GDP"
- "Implement a carbon tax by 2026"
- "Create a joint procurement system for military equipment"

What it is NOT:

- Vague aspirations without specific actions
- Proposals made by other speakers
- Background context about existing policies

4. Coding Criteria

4.1 When to Give a Thumbs Up

Award a thumbs up when:

- The extraction accurately reflects content attributed to the specified speaker

- The information is correctly categorised (e.g., issues appear in the Issue summary, proposals appear in the Proposal summary)
- All relevant information for that category is captured (completeness)
- No information from other speakers is included
- The summary is clear and concise
- **Special case**: The field is left empty correctly (or shows "No [type] summary extracted") when no content attributed to the speaker could be extracted from the summary

4.2 When to Give a Thumbs Down

Award a thumbs down when:

- Information from other speakers is included
- Relevant information from the specified speaker is missing (incompleteness)
- Information is miscategorised (e.g., a Proposal appears in Issue summary)
- The extraction is vague or unclear
- The extraction distorts or misrepresents what the speaker said
- **Special case**: The field contains content when the speaker made no relevant contributions in that category

4.3 Notes Field Guidance

Provide notes when there is something unusual or difficult about the coding decision. Your notes should:

- Specify what is unusual (e.g., "Includes Jessika Roswall's argument, not the specified speaker's")
- Indicate what is missing (e.g., "Omits speaker's proposal about joint procurement")
- Explain miscategorisation (e.g., "The content in Issue summary is actually a Position")
- Be brief but specific enough for review

5. Examples

Example 1: Correct Extraction (All Thumbs Up)

Speaker: Guy Verhofstadt

Debate excerpt: "Guy Verhofstadt emphasises that fragmented defence systems are costly and inefficient. He argues that Europe must develop independent military capabilities. He proposes creating a European Defence Union by 2027."

LLM Output:

- Issue Summary: "Fragmented defence systems are costly and inefficient"
- Position Summary: "Europe must develop independent military capabilities"
- Argument Summary: "Current fragmented systems lead to waste and reduced effectiveness"

- Proposal Summary: "Create a European Defence Union by 2027"

Coding: ✓ ✓ ✓ ✓ (all thumbs up)

Example 2: Incorrect Attribution (Thumbs Down)

Speaker: Nathalie Loiseau

Debate excerpt: "Nathalie Loiseau supports diplomatic solutions. However, Jessika Roswall argues that military deterrence is necessary given recent aggression."

LLM Output:

- Position Summary: "Supports diplomatic solutions but acknowledges military deterrence is necessary"

Coding: ✗ (thumbs down) **Notes**: "Includes Jessika Roswall's position on military deterrence, not Loiseau's"

Example 3: Miscategorisation (Thumbs Down)

Speaker: Guy Verhofstadt

LLM Output:

- Issue Summary: "Proposes establishing a European Defence Community"

Coding: ✗ (thumbs down) **Notes**: "This is a proposal, not an issue. It should appear in Proposal Summary instead"

Example 4: Empty String Correct (Thumbs Up)

Speaker: Maria Schmidt

Debate excerpt: "The debate focused on military spending. Nathalie Loiseau and Guy Verhofstadt presented extensive proposals."

LLM Output:

- issueSum: ""
- positionSum: ""
- argSum: ""
- propSum: ""

Coding: ✓ ✓ ✓ ✓ (all thumbs up) **Explanation**: The speaker a speech but it contained no Issues, Positions, Arguments, or Proposals, so empty strings are correct.

6. Edge Cases and Special Scenarios

6.1 Multiple Issues/Positions/Arguments/Proposals

If a speaker raises multiple issues, positions, arguments, or proposals, please make note in the textbox of the ones that showed up in the extraction.

6.2 Overlapping Categories

Sometimes a single statement might span categories:

- "We must act now [position] because climate data shows urgency [argument] by implementing a carbon tax [proposal]"

- This is acceptable if the extraction correctly separates these into appropriate categories

6.3 Implied vs Explicit Statements

✓ACCEPTABLE (Explicit or Direct)

Example 1 - Position:

- **Debate text**: "I support increased defence spending"
- **LLM extraction**: "Supports increased defence spending"
- **Verdict**: ✓ Explicitly stated

Example 2 - Position:

- **Debate text**: "We must increase our defence budget immediately"
- **LLM extraction**: "Supports increasing the defence budget"
- **Verdict**: ✓ Direct statement, acceptably paraphrased

Example 3 - Argument:

- **Debate text**: "Our current weapons systems are fragmented across 27 countries, leading to duplication and waste. This costs European taxpayers billions."
- **LLM extraction**: "Fragmented weapons systems create inefficiency and financial waste"
- **Verdict**: ✓ Directly stated reasoning, properly condensed

Example 4 - Proposal:

- **Debate text**: "I call on the Commission to table legislation for a European Defence Union before the end of this year"
- **LLM extraction**: "Proposes the Commission introduce European Defence Union legislation by year-end"
- **Verdict**: ✓ Clear, explicit proposal

✗ NOT ACCEPTABLE (Too Much Inference Required)

Example 1 - Over-inference:

- **Debate text**: "We cannot continue with inadequate budgets"
- **LLM extraction**: "Supports a 2% GDP defence spending target"
- **Verdict**: ✗ The speaker mentions inadequate budgets but doesn't specify what level would be adequate. The LLM has inferred a specific policy position not stated.

Example 2 - Over-inference:

- **Debate text**: "The United States has been a valuable partner, but circumstances are changing"
- **LLM extraction**: "Europe should reduce dependence on the United States"
- **Verdict**: ✗ Too much interpretive leap. The speaker notes change but doesn't explicitly advocate for reduced dependence.

Example 3 - Over-inference:

- **Debate text**: "Many member states have expressed concerns about the current approach"
- **LLM extraction**: "Opposes the current policy"
- **Verdict**: ✗ The speaker reports others' concerns but doesn't explicitly state their own position.

Example 4 - Attribution confusion:

- **Debate text**: "Some argue we need more military spending, but we must consider diplomatic solutions first"
- **LLM extraction**: "Supports increased military spending"
- **Verdict**: ✗ The speaker is presenting others' arguments and then countering them. The LLM has misattributed the position.

BORDERLINE CASES (Use Judgement)

Example 1:

- **Debate text**: "This fragmented approach is simply not working. We need coordination, we need efficiency, we need a unified system"
- **LLM extraction**: "Current fragmented defence approach is ineffective; supports creating a unified defence system"
- **Verdict**: ✓ Acceptable - while "unified system" paraphrases "coordination/efficiency/unified system", it's a reasonable synthesis of explicitly stated views

Example 2:

- **Debate text**: "When we look at the threats we face – cyber attacks, hybrid warfare, territorial aggression – it's clear our current capabilities are insufficient"
- **LLM extraction**: "Current defence capabilities are insufficient to address modern threats"
- **Verdict**: ✓ Acceptable - the speaker explicitly states capabilities are insufficient

General Rule: If you need to ask "Is the LLM reading between the lines here?" and the answer is more than a minimal "yes," code it as thumbs down. The extraction should reflect what the speaker actually said, not what they might have meant.

6.4 Attribution Ambiguity

If the debate summary is unclear about attribution, give the benefit of doubt to the LLM unless there's clear evidence of misattribution.

6.5 Paraphrasing

The LLM output doesn't need to use exact wording. Accurate paraphrasing is acceptable if meaning is preserved.

7. Quality Control Tips

1. **Read thoroughly**: Always read the full debate summary before coding

2. **Check attribution carefully**: Verify that extracted content truly belongs to the specified speaker
3. **Consider completeness**: Ask yourself, "Did the LLM capture the main points?"
4. **Watch for category mixing**: Ensure Issues, Positions, Arguments, and Proposals aren't confused
5. **Be consistent**: Apply the same standards across all validation tasks
6. **When in doubt**: If borderline, lean towards thumbs up if the extraction is substantially correct and make a note of origins of doubt
7. **Document concerns**: Use notes liberally to explain decisions

8. Common Mistakes to Watch For

- **Cross-speaker contamination**: Content from other speakers appearing in the extraction
- **Category confusion**: Positions appearing in Issue summary, proposals appearing in Argument summary, etc.
- **Incomplete extraction**: Missing significant contributions from the speaker. Make a note of such cases in the open text box
- **Overextraction**: Including tangential points not central to the speaker's contribution
- **False positives**: Extracting content when the speaker made no relevant contributions

9. Final Checklist

Before finalising your coding, ask:

- Is all extracted content attributed to the correct speaker?
- Is each piece of information in the correct category?
- Are the main points captured?
- Are there any obvious omissions?
- Have I documented my reasoning for any tricky decisions?

Remember: Your validation helps improve the accuracy of automated extraction. Be thorough, consistent, and don't hesitate to provide detailed notes when issues arise.