Can an LLM Replace a Trained **Human Coder**?

Measuring parliamentary deliberation quality at scale

AUTHOR
Feiyang Liu

SUPERVISION
Michael Grauwde
Resp. Prof. Dr. W. Brinkman

RESEARCH QUESTION *Can a **multi-model LLM ensemble** apply a theory-grounded three-dimension **Discourse Quality Index** sub-codebook reliably enough to substitute for a trained human coder?*

I. Aim

Build a reusable protocol for when an LLM can be trusted to scale a theory-grounded discourse-quality instrument.

- Defend a 3-dimension sub-codebook from deliberation theory.
- Benchmark an LLM-as-judge pipeline against trained humans.
- Diagnose where and why the LLM diverges from humans.

Result: on justification & recognition the ensemble can scale DQI coding to national-corpus size.

II. The Scaling Problem

The Discourse Quality Index is the standard tool — but does not scale.

- Each act needs two trained humans.
- Throughput ~5 acts per coder-hour.
- 200 Hansard sessions → ~1,000 person-hours.

LLMs apply rubrics cheaply — but trustworthiness must be proven

III. Three Sub-Questions

- **SQ1 · theoretical.**

Which DQI dimensions are justified for public-safety debate?

- **SQ2 · empirical.**

How reliable is the LLM ensemble vs. humans on Gwet's AC1?

- **SQ3 · diagnostic.**

Where and why does the LLM diverge from humans?

SQ1 from literature. SQ2 & SQ3 from the 200-act benchmark.

IV. Method Pipeline

① Corpus
200 UK Hansard public-safety acts · 4 sub-domains · 50 each · stratified by length · 20–200 words.

▼

② Parallel coding · Human coders + LLM ensemble
HU1 & HU2 (trained). 4-model ensemble: sonnet · haiku · llama · qwen · self-consistency k=5.

▼

③ Reliability matrix · 3 × 3
Gwet's AC1 (primary) · Cohen's κ_w · Krippendorff's α · 1,000× bootstrap 95 % CI.

▼

④ Substitution decision · pre-registered
Equivalence: δ = AC1(LLM,HU1) − AC1(HU1,HU2); substitutes if 95 % CI lies inside ±0.10 band.

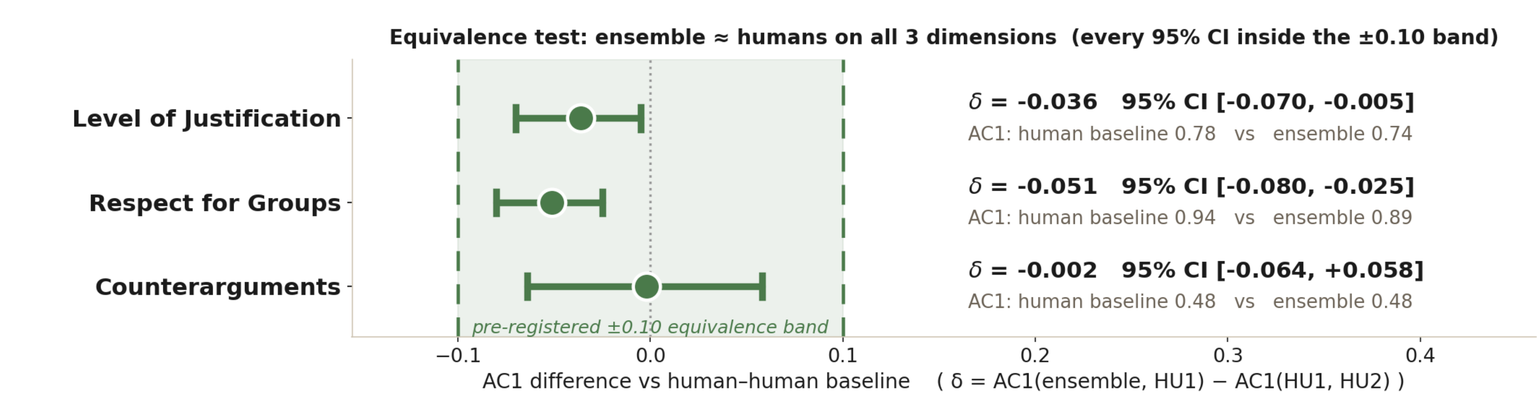
Cut-offs cited from Dunivin (2024) & Chew et al. (2023)

V. Sub-Codebook

Six DQI dimensions exist. Three are kept; three dropped.

Level of Justification 0–3	Reason-giving is foundational.
Respect for Groups 0–2	Mutual recognition.
Counterarguments 0–2	Separates deliberation from debate.
Participation 0–2	System-level; fixed by Hansard procedure.
Content of Justification 0–3	Theory-contested
Constructive Politics 0–3	Floor effect in adversarial Hansard debate.

VI. Results · Ensemble vs Trained Coders



VII. Conclusion

✓ LLM ≈ humans on all 3 dims	YES
✓ LLM-only: Justification + Respect	YES
✓ Counterarguments: as 2nd coder	YES
→ Counterarguments: as sole coder	NO