Scaling Deliberative-Quality Measurement with Large Language Models

Validating a Four-Model LLM Ensemble as a Coder for a Theory-Justified Three-Dimension Discourse Quality Index Sub-Codebook on UK House of Commons Debate

Feiyang Liu[1]

Responsible Professor: Dr. William Brinkman[1]
Supervisor: Michael Grauwde[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 17, 2026

Name of the student: Feiyang Liu
Final project course: CSE3000 Research Project
Thesis committee: Dr. William Brinkman, Michael Grauwde, [Examiner TBA]

Abstract

Evaluating the quality of public deliberation is a prerequisite for governing it on evidence rather than impression, yet the bottleneck is measurement. The Discourse Quality Index (DQI), the standard instrument in the field, requires trained human coders — slow, costly, and applied inconsistently across research teams — which means most empirical deliberation studies contain only a few thousand speeches. Scaling that measurement is the problem this paper addresses. Large language models (LLMs) can apply a structured rubric rapidly and uniformly, so the central question is whether an LLM coder is reliable enough on theory-grounded deliberative constructs to stand in for a trained human and raise the data-size ceiling. We make two contributions. First, we construct a theory-justified three-dimension DQI sub-codebook — Level of Justification, Respect for Groups, and Counterarguments — defending each inclusion and exclusion from deliberative-democracy theory. Second, we benchmark a four-model LLM-as-judge ensemble against two trained coders on 200 UK House of Commons public-safety debate acts. We report Gwet's linear-weighted AC1: trained coders reach AC1 = 0.78, 0.94, and 0.48 on the three dimensions, and the ensemble reaches 0.74, 0.89, and 0.48 against the lead coder. We also tested whether an LLM can substitute for a human coder, framed as an equivalence test: whether the paired difference in agreement falls within a ±0.10 band of the human–human baseline, judged by where its confidence interval lies rather than by a test against zero. The difference falls within 0.05 of the baseline on every dimension, and its confidence interval lies inside the band on all three. The results suggest that on justification and recognition the ensemble can serve as reliable independent measurement, and that on Counterarguments it can serve as a second coder in a doubly-coded design — where the binding constraint appears to be the codebook anchor language rather than the model.

Keywords: deliberative-quality measurement, Discourse Quality Index, LLM-as-judge, automated annotation, inter-rater reliability, Gwet's AC1, deliberative democracy.

1 Introduction

In the deliberative-democracy tradition, democratic deliberation is the exchange of reasons through which participants justify their positions to one another and work toward collective decisions [6,7,18]. A decision is legitimate when it can be defended with reasons that others could accept on reflection, rather than imposed by power or persuasion alone [6,7]. As deliberation is increasingly conducted at scale: in online consultations and large citizens' assemblies [53], and around AI-assisted decision systems such as predictive policing, evaluating such settings on the basis of evidence — a measured account of how well participants actually reason, engage opposing views, and treat affected groups — rather than on the basis of impression requires assessing the deliberation at a volume that manual reading cannot reach. Measuring deliberative quality is therefore a prerequisite for evidence-based evaluation of mediated public deliberation, and it is where the Discourse Quality Index (DQI), face a bottleneck.

The DQI [46] turns Habermasian discourse ethics into ordinal scores on six dimensions of a single speech act. It has become a widely used measure in empirical deliberation research, applied and adapted in multiple parliamentary and deliberative settings [27,41]. The barrier is not the instrument but the labour required to apply it. Coding requires trained human coders working independently, and double coding is needed to assess inter-rater reliability [46]; in addition, different research teams operationalise the same constructs differently [15]. At the scale of thousands or tens of thousands of speech acts, this creates a substantial manual-coding burden. The result is that DQI studies are confined to dataset of a few thousand acts, that comparison across teams is fragile, and that the actors who most frequently uses the measure — policymakers commissioning a

consultation, or participants who want feedback on the quality of a deliberation while it is still running — cannot obtain it at the speed or scale they need. Removing this labour barrier, without abandoning the theory-grounded measurement the DQI provides, is the problem this paper addresses.

Large language models (LLMs) may offer a route past this barrier, because they can apply a fixed rubric to thousands of acts rapidly and at uniform cost. The LLM-as-judge paradigm [54] and a growing computational-social-science literature [14, 55] report that recent instruction-tuned models can apply structured rubrics to subjective political text at agreement comparable to human coders on some tasks. Substituting an LLM coder for a human one is thus a scalable method — but its reliability cannot be assumed. Replication work shows that LLM codes diverge from human gold labels in task-specific, prompt-sensitive ways, and that the divergence tends to be largest on exactly the subjective and theoretical constructs the DQI targets [23, 40]. Whether an LLM meets the reliability bar for a given deliberative construct is therefore an empirical question, and it is the question this paper addresses.

Rather than work with all six DQI dimensions, we reduce the instrument to three — Level of Justification, Respect for Groups, and Counterarguments — which the deliberative-democracy literature treats as the micro-level core of deliberative quality, and set the other three aside as system-level, normatively contested, or floor-effected at the single-act level. We give the theoretical case for this selection in Section 2 and the full per-dimension warrants in Appendix B.1. We had fixed these three dimensions, and the three exclusions, in the pre-registration before coding any act, so the reduction is part of the study design rather than a choice made after seeing which dimensions happened to work. We then report reliability for each retained dimension separately, including the weakest, since showing where the substitution holds and where it breaks down is part of what the study contributes.

This gives the central research question:

> *Can a multi-model LLM ensemble apply a theory-grounded three-dimension Discourse Quality Index sub-codebook to deliberation on public-safety policy reliably enough to substitute for a trained human coder?*

The question is about public-safety deliberation as a domain, not about any single dataset; we operationalise it in the Method (Section 3) using UK House of Commons public-safety debate as the corpus. The substitution test (Section 3.4) operationalises "reliably enough" as an equivalence criterion: LLM–human agreement must fall within a fixed band of the human–human baseline on each dimension, judged by where the confidence interval lies rather than by a test of zero difference. The paper makes two contributions toward this question. The first is a theory-justified three-dimension sub-codebook with explicit inclusion criteria for Level of Justification, Respect for Groups, and Counterarguments, and exclusion criteria for the other three DQI dimensions. The second is a four-model LLM-as-judge pipeline, benchmarked against two trained coders on a 200-act UK House of Commons calibration set, with reliability assessed using Gwet's AC1 and the equivalence test.

The remainder of the paper reviews related work (Section 2); presents the sub-codebook, the LLM pipeline, the corpus, and the pre-registered analysis plan (Section 3); reports the reliability results (Section 4); discusses possibility of scaling deliberative-quality measurement using llm and how it would fail (Section 5); and addresses responsible research (Section 6), before concluding (Section 7).

2 Related Work

2.1 LLM annotation of subjective political text

This work builds on a body of evidence that recent LLMs can reach human-comparable agreement on political-text annotation. GPT-4 has been reported to match or exceed expert and crowd coders on party-affiliation classification of tweets [48], and expert agreement reaching 95% on short annotations has been found to fall on longer articles [20]. At the harder end, Ziems et al. [55] benchmark thirteen models across twenty-five tasks and find subjective, value-laden constructs — where DQI is employed — among the weakest for llm annotation, and Dunivin [9] reaches human-equivalent agreement on only three of nine socio-historical codes under a single closed model scored with Cohen's kappa. Read together, these studies suggest that agreement on long-form, value-based constructs is real but uneven, which is why the present design draws on a long-form speech-act register, a chance-corrected statistic, a multi-model ensemble rather than one closed model, per-dimension reporting, and a pre-registered plan.

The same literature also supplies a set of validation practices that this design adopts. Prior work treats mandatory human validation as the baseline [39, 40]; uses repeated sampling to control the variance of a single prompt [43] (Section 3.2); and adds open-weights comparators as a safeguard against closed-model reproducibility concerns [23, 38]. It has further converged on checking that logged rationales are faithful rather than post-hoc [25, 49], examining possible temporal sensitivity in annotation results [1], and reporting robustness to model and prompt choices [1, 49]. We carry each of these practices into the pipeline described in Section 3.

2.2 Automated assessment of deliberation quality

The closest deliberation automation work is AQuA [3], which fine-tunes BERT-adapter models on a 20-indicator codebook over German online comments and aggregates the indicators into a single deliberativeness score. AQuA is the natural foundation to build on: it demonstrates that automated deliberative-quality measurement is feasible and that a theory-derived codebook can be scaled. The present study extends it along three lines. It uses instruction-tuned generative LLMs, which can apply a written rubric zero-shot without the labelled fine-tuning data an adapter model requires; it keeps the theory-grounded dimensions separate rather than collapsing them into one score, because prior human-coding work shows deliberative-quality variance concentrates by dimension and register [27], so a single pooled score discards information this design preserves; and it codes long-form parliamentary debate rather than short online comments. Related work has also compared debate-quality indices in social-media discussions [44].

Table 1: How this study relates to the closest prior work on LLM and automated annotation of subjective political and argumentative text.

Study	**Approach**	**Construct (register)**	**Reliability metric**	**Human validation**
Gilardi et al. [14]	Zero-shot ChatGPT	4 annotation tasks (tweets, news)	% agreement vs. crowd	MTurk crowd
Törnberg [48]	GPT-4, zero-shot	Party affiliation (tweets)	Accuracy	Expert & crowd
Dunivin [9]	GPT-4 + chain-of-thought	9 socio-historical codes (paragraphs)	Cohen's κ (human-equiv. 3/9)	Expert gold labels

Mirzakhmedova et al. [34]	LLM annotators	15-dim argument quality (argumentative text)	Agreement, pooled across dims	Expert; not pre-registered
Behrendt et al. [3] — AQuA	Fine-tuned BERT adapters	20 indicators → 1 score (German online comments)	Correlation with scores	Expert + non-expert
This work	Instruction-tuned 4-LLM ensemble + self-consistency	3-dim DQI sub-codebook (UK parliamentary debate)	Gwet's AC1 (+ κ, α), paired δ	2 trained coders, pre-registered

2.3 From the DQI to three dimensions

The DQI scores six dimensions of a single speech act [46], but not all six are equally suited to single-act coding, and this is where the three retained dimensions come from. Three are widely treated as the micro-level core of deliberative quality: reason-giving and mutual recognition are presented as the minimum conditions of deliberative legitimacy [6, 18, 37], and engagement with counterarguments is what most clearly separates deliberation from parallel monologue [2, 36]. The other three fit less well at the level of single acts: Participation and Constructive Politics are better understood as system-level properties [16, 31], and Content of Justification depends on a value preference that remains contested [32]. In directed content analysis, keeping a category means showing it is warranted by prior theory [21]; on that standard the three we retain — Level of Justification, Respect for Groups, and Counterarguments — are easier to defend than the three we set aside. There is also a practical reason to simplify: a review of 67 studies and 123 indicators finds low criterion validity across many DQI-style operationalisations, pointing to the number and abstraction of the criteria as a recurrent source of inconsistency [15], so a shorter, more concrete instrument should be easier to apply reliably. We carry this selection into the sub-codebook of Section 3, where the dimensions are operationalised with anchors matched to the public-safety register; the full per-dimension warrants are in Appendix B.1.

3 Method

Based on DQI, we come up with a written sub-codebook, have independent raters apply it to a common set of items, and quantify how closely raters agree. The novelty is in the raters, one of which is an LLM ensemble rather than a human. We choose this design because it isolates the question of interest — can an LLM reproduce a trained human's application of a theory-grounded rubric — from common practices in model fine-tuning or task-specific supervision: the ensemble sees only the same written anchors a human coder is trained on, so agreement (or its absence) can be attributed to the rubric and the model, not to the training data. Reliability is assessed with chance-corrected agreement statistics and an equivalence test against a human–human baseline [22].

Concretely, this section presents the three-dimension sub-codebook (Section 3.1), the LLM coding pipeline (Section 3.2), the dataset and human-coder protocol (Section 3.3), and the pre-registered analysis plan (Section 3.4). Figure 1 shows the design overview. The supplementary reproducibility bundle (codebook, dataset, prompts, model-parameter log, helper code, and README) is archived on 4TU.ResearchData and is also available on GitHub at https://github.com/feiyangliu2023/ResearchProjectDelftReproducibility.

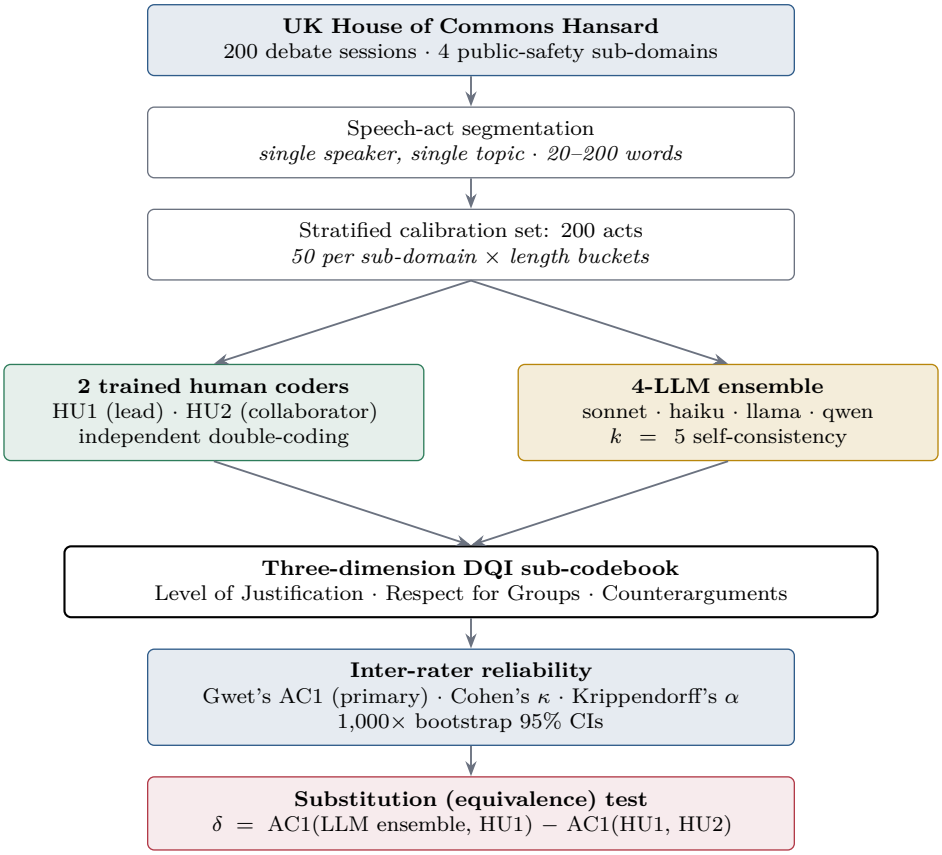

Figure 1: Study design and analysis pipeline. A stratified 200-act calibration set drawn from UK House of Commons public-safety debate is coded independently by two trained humans and a four-model LLM ensemble on three DQI dimensions; agreement is measured with Gwet's AC1 and the substitution claim is tested with a paired δ.

3.1 A three-dimension DQI sub-codebook

The DQI scores individual speech acts on six dimensions [46]. We retain three — Level of Justification, Respect for Groups, and Counterarguments — on the theoretical grounds set out in Section 2.3, and operationalise them here with anchors matched to the public-safety register.

Included dimensions. Level of Justification (0–3) captures reason-giving, the foundational claim of deliberative democracy [6, 18], and carries most of the deliberative variance in prior empirical work [13, 41]. Respect for Groups (0–2) is the inclusion principle [12, 35, 53] and is especially important in the public-safety context, where predictive-policing, recidivism, and biometric-surveillance systems disproportionately affect minority groups [10]. Counterarguments (0–2) is the marker that most clearly separates deliberation from debate [2], and has direct precedent in the reduced three-dimension DQI of Steiner et al. [47]. All three are demanding LLM targets, since coding it requires identifying the opposed position and judging whether the speaker engages its substance. Scale anchors are in Table 2; full inclusion and exclusion criteria follow the MacQueen et al. [28] six-component codebook structure.

Excluded dimensions. The three remaining dimensions were not coded. Participation is a procedural, system-level property rather than a property of an individual speech act, and is near time-invariant once turn allocation is controlled by a presiding officer. Content of Justification places common-good appeals highest, and this is contested in the theory by Mansbridge et al. [32] and shown to be unstable across contexts [29]. Constructive Politics is liable to floor effects in adversarial, multi-party debate [33]. Including any of the three would therefore import a contested normative ranking, a system-level construct, or a structurally floor-effected one into a single-act instrument. Figure 2 summarises the selection; the full per-dimension criteria are in Appendix B.1.

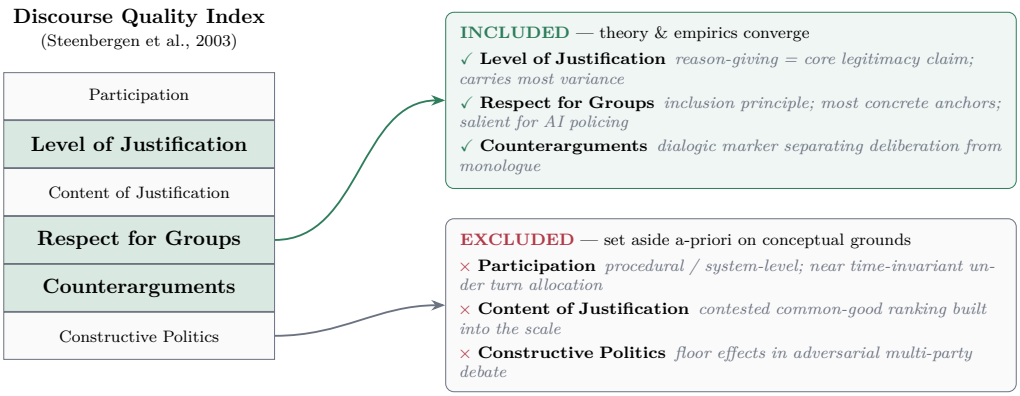

Figure 2: From the six-dimension DQI to the theory-justified three-dimension sub-codebook. Three dimensions are retained where deliberative-democracy theory and prior empirical work converge; three are set aside a-priori on conceptual and construct-validity grounds — a system-level construct, a contested normative ranking, and a floor-effected dimension. The selection was pre-registered before any reliability analysis.

Table 2: Scale anchors for the three retained DQI dimensions in the public-safety register.

Dimension	Score	Anchor (Hansard / public-safety example)
Level of Justification	0	Bare assertion. "We must ban facial recognition in public spaces."

	1	Inferior justification: a single weak premise. "... because it is wrong."
	2	Qualified justification: a recognisable argument with an identifiable premise.
	3	Sophisticated justification: a multi-premise argument with explicit linkages.
Respect for Groups	0	Explicit derogation of a social group, including stereotyping.
	1	Neutral reference: a group named without evaluative framing.
	2	Explicit affirmation of a group's dignity, standing, or contribution.
Counterarguments	0	No engagement with opposing positions.
	1	Acknowledged but not engaged: an opposing position is named or quoted but dismissed without substantive reply.
	2	Engages substantively with the opposing position's reasons.

3.2 LLM coding pipeline

Architecture. The pipeline scores one speech act on one dimension at a time, producing an integer score and a written rationale, following the G-Eval architecture [26]: a first prompt derives a sequence of evaluation steps from the codebook entry, and a second prompt applies those steps to the act and returns structured JSON of the form {"score": int, "rationale": str}. This two-stage prompting design [51] makes the evaluation procedure explicit before scoring. Robustness to known LLM-judge biases and pipeline choices [54] is assessed separately through the sensitivity analyses described below. Two extensions are added. Self-consistency sampling [50] draws $k = 5$ samples at temperature 0.7 and takes the median score — five is the largest odd k within the roughly 12,000-call budget, odd so the median never ties. Across models the rule is a median-of-medians, with adjacent-anchor ties broken to the lower, more conservative anchor so the pipeline never inflates the contested top anchors (Section 4.1). The full pipeline is shown in Appendix Figure A.1.

Prompt design. Prompts follow the failure-mode classification of Zheng et al. [54] and score one dimension at a time on semantic content, with the rationale required before the score; the full prompt text for every dimension is reproduced verbatim in Appendix A.

Models and reproducibility. Following multi-model recommendations [23, 38], the ensemble uses four instruction-tuned models — two closed-weights (claude-sonnet-4.6, claude-haiku-4.5) and two open-weights (llama-3.3-70b-instruct, qwen-2.5-72b-instruct). Temperature is 0.7 for sampling, seeds are fixed where the API exposes them, and every call is logged with timestamp and content hash for independent replication; the full parameter manifest is in the bundle.

3.3 Corpus, calibration subset, and metrics

Corpus. The corpus is 200 UK House of Commons debate sessions scraped from hansard.parliament.uk and filtered to four public-safety sub-domains: crime and policing; emergency services and public-protection technology; surveillance, biometric data, and

civil liberties; and terrorism and counter-terrorism. Following Steenbergen et al. [46], the unit of analysis is an individual speech contribution. For the present study, we operationalise speech acts as contiguous contributions by one speaker on one topic, capped at 200 words and split at within-turn topic shifts; these word limits and the topic-shift splitting are study-specific operational decisions. Contributions under 20 words are excluded as insufficient for reliable coding.

Calibration subset and coders. Two trained human coders — denoted HU1 and HU2 — code each act independently on all three dimensions: Coder 1 (HU1) is the lead author and Coder 2 (HU2) a trained collaborator. The human reliability baseline is therefore an agreement between two independent human raters applying the same written codebook to the same 200 acts: HU1 and HU2 code blind to each other, and inter-rater reliability is the agreement between their two independent code sets. Both coders complete a two-session training phase on practice transcripts outside the reliability set, after which coding is fully independent until all data is coded. The calibration subset contains 200 acts. Uncertainty around the reliability estimates is quantified using 1,000-resample percentile-bootstrap 95% confidence intervals at the act level.

Reliability metrics. Cohen's weighted kappa is conventional in analysis of previous research, but on highly skewed marginals it collapses toward zero even when raters agree on most items — the "kappa paradox" [11,52]. Because every dimension here concentrates most acts on a single anchor (Respect for Groups holds roughly 90% at the centre), the headline statistic is Gwet's linear-weighted AC1 [17], which replaces the chance term that skewed marginals inflate with one that stays stable under single-category dominance; Cohen's weighted kappa is reported for back-compatibility and Krippendorff's alpha as a secondary check. The pre-registration had named Cohen's weighted kappa as the primary metric, with AC1 as a guard against the high-prevalence paradox; because that paradox did appear in these marginals, we use AC1 as the main metric and keep kappa alongside it. AC1 is reported per dimension and per coder pair (HU1–HU2, LLM–HU1, LLM–HU2) with 1,000-resample percentile-bootstrap 95% CIs (seed 20260519).

To test whether LLM can be good substitute to human coder. We use a two-sided equivalence (TOST) test rather than a one-sided non-inferiority test because, for a measurement instrument, the LLM must be neither materially worse than the human baseline nor materially more agreeing: an agreement that overshoots the human–human band would point to anchor-gaming or mode collapse rather than to faithful coding, so deviations in either direction are disqualifying and a two-sided band is the appropriate criterion. The quantity tested is the paired difference in agreement, $\delta = \text{AC1}(\text{LLM ensemble, HU1}) - \text{AC1}(\text{HU1, HU2})$, computed per dimension at the act level. We call the ensemble substitutable on a dimension when the 95% confidence interval for δ lies entirely inside the ± 0.10 band. The decision is read from where that interval sits, not from whether δ differs from zero: a small difference can be detectable and still equivalent, and a non-significant difference is not on its own evidence of equivalence. We anchor δ on HU1; because HU1 wrote the codebook, the ensemble — which sees only the written anchors — should track the other coder, HU2, more closely.

3.4 Analysis Plan

The full analysis plan, the three operational hypotheses (labelled RH1–RH3 to keep them distinct from the human coders HU1 and HU2), the dimension selection, and the codebook-stability rule are time-stamped (2026-05-19) in the 4TU.ResearchData record, alongside the reproducibility bundle. The registration was filed at the design stage: after the codebook and corpus-construction procedure were fixed but before the calibration codes were produced and before any reliability statistic was computed. The exclusion of the three DQI dimensions is part of that timestamped registration.

In brief, the substitutability condition is that the paired difference in agreement falls within the equivalence band $[-0.10, +0.10]$ of the human–human baseline on each dimension. The ± 0.10 bound was preregistered as the study-defined smallest effect size of interest for the difference between LLM–human and human–human agreement. It represents the maximum difference the study considered practically acceptable for coder substitution. Sensitivity to alternative bounds should be considered when interpreting the result. Two minor execution deviations are detailed in Appendix B.8.

Post-hoc validation checks. Beyond the tests, three additional checks assess whether the result is an artefact of the rationales being decoupled from the scores, varies across debate years, or depends on incidental pipeline choices. A reasoning-faithfulness audit asks whether each call's written rationale concludes the anchor it actually emits, automated over all calls and validated by hand on a stratified sample, following the manual reasoning checks of Lin and Zhang [25] and the post-hoc-rationalisation of Törnberg [49]. A temporal robustness check compares agreement across debate years, including acts from 2020, following the train-versus-test framing of Abraham et al. [1]: if agreement were driven by exposure to older text rather than the rubric, it would decline on the most recent acts, which sit at or beyond the models' training windows. This check can test temporal stability but cannot by itself establish training-data contamination. A robustness check re-runs the substitution δ under leave-one-model-out ensembles and three alternative aggregation rules [39,49]. All three are computed on the released per-call codes; the procedures and full tables are in Appendix B.9–B.11.

4 Results

All three coders applied the Section 3 codebook to the same 200 acts; the LLM ensemble coded at k = 5 self-consistency with a median-of-medians headline. The results are reported in four steps that follow the research question: how the coders distribute labels (4.1), pairwise reliability (4.2), the substitution test that answers the question (4.3), and where the residual disagreement falls (4.4). AC1 is the primary statistic. Per-model, inter-model, and variance analyses are in Appendix B.

4.1 Label distributions

Coders agree on the central anchors but diverge on the top anchors that distinguish high-quality deliberation from generic political speech (Figure 3). On Level of Justification the qualified anchor (LJ = 2) carries 44–78% of acts under every coder, but the sophisticated anchor (LJ = 3) is reached on 9.5% of acts under HU1 versus 0.5% under HU2. On Respect for Groups the centre anchor holds 76–94% of acts everywhere. On Counterarguments, HU1 splits acts almost evenly between CA = 0 (50.5%) and CA = 2 (47.5%), whereas HU2 places 93% at CA = 0 and 1% at CA = 2. The ensemble's distribution sits closer to HU2's conservative reading on both LJ = 3 and CA = 2, which is what drives the asymmetry in the reliability matrix below. Full per-coder counts are in Appendix Table B.1.

4.2 Pairwise reliability

The human baseline is the agreement between the two human coders, HU1 and HU2, and the LLM ensemble is then compared against each of them in turn. Table 3 reports the headline AC1 matrix, with Cohen's weighted kappa for back-compatibility. The human–human baseline is substantial on Level of Justification (AC1 = 0.78), almost perfect on Respect for Groups (0.94), and only moderate on Counterarguments (0.48). The kappa paradox is visible in the table: on Respect for Groups, where about 90% of

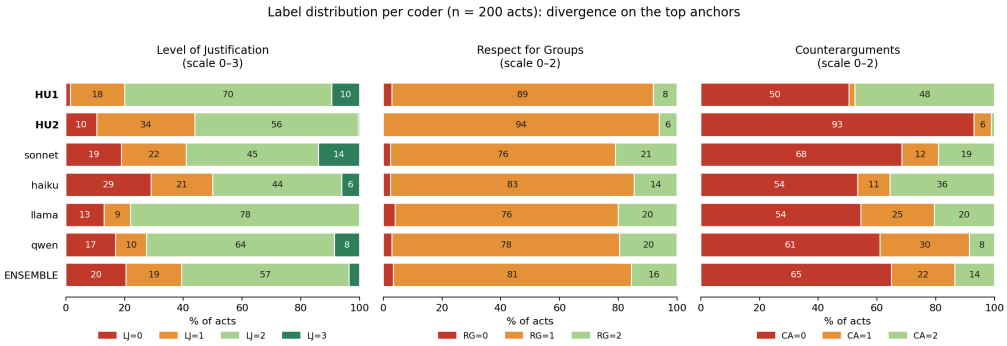

Figure 3: Label distribution per coder (n = 200). Coders agree on the central anchors but diverge on the top anchors that distinguish high-quality deliberation — most visibly on Counterarguments, where HU1 reaches CA = 2 on 48% of acts and HU2 on 1%.

acts sit at one anchor, AC1 is in the 0.89–0.94 band across all three pairs while Cohen's weighted kappa collapses to 0.23–0.40 (a side-by-side plot is in Appendix Figure B.1). Against the baseline, the LLM ensemble is uniformly close to HU1 (ΔAC1 = −0.04 on LJ, −0.05 on RG, 0.00 on CA) and uniformly above the baseline against HU2 — the asymmetry follows from the distributional pattern of Section 4.1, where the ensemble and HU2 share a conservative top-anchor reading.

Table 3: Pairwise inter-rater reliability on the 200-act subsample. Headline AC1 (bold) with Cohen's weighted kappa; CIs are 1,000-resample percentile bootstraps, seed 20260519.

Pair	Dim	$\mathbf{AC1_{vv}}$ (95% CI)	$\boldsymbol{\kappa_{vv}}$ (95% CI)
HU1–HU2	LJ	**0.78** (0.738, 0.810)	0.12 (0.032, 0.206)
HU1–HU2	RG	**0.94** (0.913, 0.961)	0.25 (0.046, 0.452)
HU1–HU2	CA	**0.48** (0.393, 0.558)	0.05 (0.015, 0.100)
HU1–ENS	LJ	**0.74** (0.698, 0.783)	0.10 (0.022, 0.188)
HU1–ENS	RG	**0.89** (0.854, 0.918)	0.23 (0.076, 0.373)
HU1–ENS	CA	**0.48** (0.397, 0.548)	0.10 (−0.006, 0.199)
HU2–ENS	LJ	**0.87** (0.842, 0.897)	0.57 (0.470, 0.658)
HU2–ENS	RG	**0.93** (0.898, 0.952)	0.40 (0.213, 0.562)
HU2–ENS	CA	**0.76** (0.707, 0.818)	0.17 (0.094, 0.265)

4.3 The substitution test

We test substitutability with the equivalence test on the paired difference δ (Table 4, Figure 4). The rule is simple: the ensemble counts as substitutable on a dimension if the 95% confidence interval for δ falls inside the ± 0.10 band, whether or not δ itself differs from zero. On Level of Justification and Respect for Groups the interval is inside the band but excludes zero, so the ensemble agrees with HU1 a little less than the two humans agree with each other — a real but small gap that stays within the bound. On Counterarguments the interval is inside the band and also contains zero, so we cannot detect any difference from the human baseline — though that baseline is itself only moderate (AC1 = 0.48).

Table 4: Equivalence test on the paired δ = AC1(LLM ensemble, HU1) − AC1(HU1, HU2) per dimension (point estimate bold). The ensemble is substitutable on a dimension when the 95% CI for δ lies inside the ±0.10 band. CIs are paired bootstraps at the act level.

Dim	δ AC1 (95% CI)	95% CI within ±0.10 band?
LJ	**−0.04** (−0.070, −0.005)	yes — equivalent
RG	**−0.05** (−0.080, −0.025)	yes — equivalent
CA	**0.00** (−0.064, +0.058)	yes — equivalent

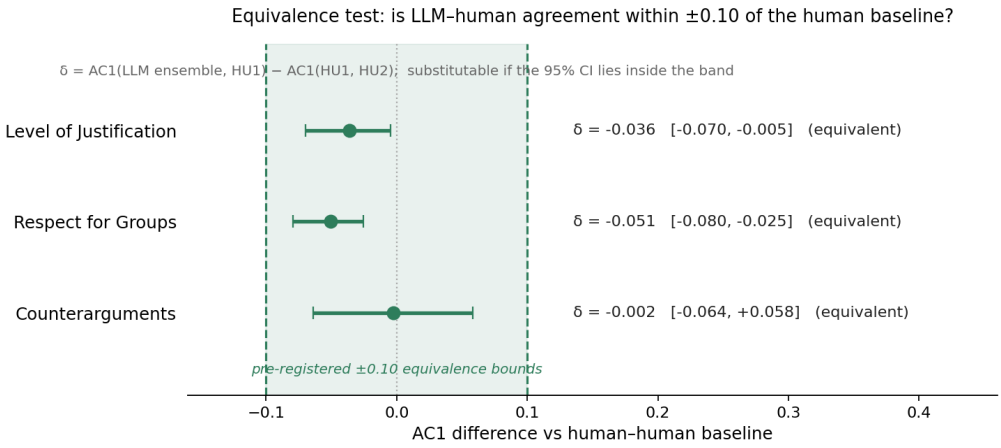

Figure 4: Equivalence test. Paired δ = AC1(LLM ensemble, HU1) − AC1(HU1, HU2) per dimension, with 95% bootstrap CIs. On all three dimensions the CI lies inside the ±0.10 equivalence band, so the equivalence criterion is met; on Counterarguments the interval additionally includes zero.

4.4 Where the disagreement falls

Decomposing each disagreement by ordinal distance separates decision-irrelevant noise from actual design problem. On Respect for Groups the ensemble's misses against HU1 are almost all off-by-one ordinal noise — gross disagreements, two or more anchors apart, are just 1.0% of acts — and on Level of Justification they are mostly off-by-one, with 15.5% gross. On Counterarguments the disagreement is categorical, not gradational: in 33.5% of acts the ensemble and HU1 sit two anchors apart (the full CA = 0↔2 distance), and 61% of their Counterarguments disagreement is gross. The ensemble does not introduce a new split — it amplifies the existing human one, siding with the conservative coder (it matches HU2 on 67% of Counterarguments acts but HU1 on only 45%). The full ordinal-distance decomposition is in Appendix Table B.4.

4.5 Robustness, faithfulness, and temporal sensitivity

Three post-hoc checks (Section 3.4) test whether the substitution result survives the standard threats the LLM-as-judge literature raises; full tables and figures are in Appendix B.9–B.11.

The rationales are faithful to the scores. Across the 11,999 calls, the LLM score matches the score concluded automatically from LLM rationale in 88.5% of calls (95% CI 87.5–89.5%). A by-hand audit of a 50-call stratified sample found no genuine contradiction between score and rationale — every automated "mismatch" in the sample was an

extraction error in which the rationale named a rejected adjacent anchor ("matching LJ = 2, but not LJ = 3"), so 88.5% is a conservative floor and the true agreement rate is likely substantially higher. The pipeline's scores are therefore not decoupled from its stated reasoning, which is the failure mode a logged rationale is meant to rule out.

Agreement does not decline for more recent acts. The corpus spans 2020–2025, and 81 of the 200 acts are from 2025, at or beyond the four models' training windows; public Hansard text from the earlier years is plausibly in pre-training. If temporal exposure rather than the rubric drove agreement, ensemble–human agreement would be highest on the oldest acts and fall on the most recent — but it is flat (Figure B.4): the Spearman correlation between debate year and ensemble–HU1 agreement is $\rho = 0.03$ (p = 0.70), and mean exact agreement is 0.58 on 2020–2024 acts versus 0.58 on 2025 acts (difference −0.003, 95% CI −0.078 to +0.072). Per-dimension AC1 against HU1 does not degrade on the post-training-window acts (LJ 0.73 → 0.75, RG 0.89 → 0.88, CA 0.46 → 0.50). The DQI labels themselves cannot have been memorised — they were produced by HU1 and HU2 for this study and were never public. The flat recency profile provides no evidence of temporal degradation in agreement, but it cannot rule out memorisation of the debate text or other training-data exposure.

The conclusion does not depend on one model or one aggregation rule. Within-model self-consistency is high: across the four models the five k-samples are unanimous on 69% (LJ), 85% (RG), and 69% (CA) of acts, with a mean modal-vote share of 0.90–0.96 and a median within-model standard deviation of zero. Re-running the substitution δ under each of the four leave-one-model-out ensembles, and under three alternative aggregation rules (mean-then-round, majority mode, and a pooled median over all twenty samples), leaves every one of the 24 paired δ point estimates inside the ± 0.10 band, and Counterarguments remains statistically indistinguishable from the human baseline under every variant (Figure B.5).

5 Discussion

5.1 What the result supports: answering the research question

The research question asked whether a four-model ensemble can apply the sub-codebook reliably enough to substitute for a trained coder and so scale deliberative-quality measurement. The answer depends on the dimension and definition of substitution in paper. We separate three senses of substitution and report which holds where (Table 5). The LLM reliably serves as a second coder in a doubly-coded design without decision-relevant change to the reliability matrix — substitution is supported on every dimension: every paired-δ CI lies within the equivalence band, and on Counterarguments the interval also includes zero. When LLM acts as sole coder with no human in the loop, it holds on Level of Justification and Respect for Groups but not on Counterarguments, where AC1 against HU1 (0.48) matches the human baseline but is too low to use in isolation. When using LLM-only scores as the actual measurement of deliberative quality — it follows for the first two dimensions, while Counterarguments scores should be human-adjudicated or excluded until the codebook is revised. No single model dominates the per-dimension test, so the median-of-medians ensemble, not any one provider, is the appropriate headline; inter-model agreement is uniformly substantial (Appendix B.4), so this cohesion reflects convergence across four independent providers rather than one model pulling the median — the reason the inter-model comparison is retained rather than cut.

Table 5: Where each sense of "LLM substitution for a human coder" is supported, by dimension (LJ = Level of Justification, RG = Respect for Groups, CA = Counterarguments).

Sense of substitution	**LJ**	**RG**	**CA**
Weak-procedural — LLM as second coder in a doubly-coded design	Supported	Supported	Supported
Strong-procedural — LLM as sole coder, no human in the loop	Supported	Supported	Not supported (AC1 = 0.48)
Strong-substantive — LLM-only score as the operational measure	Supported	Supported	Human-adjudicate or exclude

5.2 The Counterarguments constraint is the anchor, not the model

On Counterarguments the human–human baseline (AC1 = 0.48) and the ensemble against HU1 (0.48) are effectively equal, and their paired difference is equivalent to zero: the binding constraint is the codebook anchor language, not model capability. This underperforming dimension is kept and diagnosed rather than dropped: a dimension on which neither the two human coders nor the LLM reach reliable agreement is an informative negative result, because it marks the point at which the present instrument stops measuring the construct reliably. No coder, human or LLM, can reliably separate CA = 0, 1, and 2 under the codebook anchors applied to UK public-safety speech. HU1 credits any substantive engagement with an opposing position, even via reframing, as a counter-with-response (47.5% at CA = 2); HU2 reserves the top anchor for explicit rebuttals that name the counter position (1% at CA = 2); both are defensible readings of the present anchor language. Reading the per-act rationales logged for every code shows that the divergences are not noise but a small set of recurring rhetorical moves the anchor does not decide — an anticipatory or unvoiced objection, a named-but-dismissed position, or a speaker's own internal balancing — and the four models split among themselves on precisely these acts (the five-mode typology and two worked examples are in Appendix B.6). That the boundary is under-determined within a single coder, not only between coders, shows as a constraint in the instrument rather than the rater. The codebook-stability analysis (above 20% pairwise disagreement) reveals issue for Counterarguments but not for the other two dimensions, so the actionable remedy is a tightened CA = 2 anchor with worked examples drawn from the disagreement set, not further model tuning.

5.3 Implications for scaled deliberative-quality measurement

Read back onto the measurement problem that motivated the study, the result is concrete. On the dimensions where deliberative theory and prior empirics are strongest — justification and recognition — a median-of-medians LLM ensemble appears able to serve as the operational measurement at the scale of a national parliamentary corpus, easing the cost barrier that has confined DQI studies to a few thousand speeches. That makes previously impractical work tractable: longitudinal tracking of justification quality across a full parliamentary term, cross-chamber or cross-country comparison, and — most directly — quality control for the AI-generated summaries that increasingly mediate large citizens' assemblies and online consultations, where an automated DQI score can flag when a summary flattens or inflates the deliberative quality of the underlying contributions.

For a policymaker commissioning such a process, the instrument offers an auditable, theory-grounded quality signal at the data level that hand-coding cannot supply at that volume. Two boundaries keep this honest: the score is a dataset-level research signal, not a verdict on any individual speaker (Section 6), and Counterarguments must be human-adjudicated until its anchor is revised. The open question the result raises is whether a revised Counterarguments anchor lifts that dimension to the same standard.

5.4 Limitations and opportunities

The claims are bounded in normal ways, several of which double as the natural next steps. The three-dimension sub-codebook is one selection from the six-dimension DQI; the substitution claim does not transfer to the excluded dimensions, and extending the validation to them is an open opportunity for future work. The data is restricted to four UK public-safety sub-domains in English, so transfer to citizen-deliberation registers or other languages requires separately validated anchors and prompts — the most direct follow-up, since it is exactly the mediated-deliberation setting the measurement is meant to serve. The baseline is a single coder pair, and HU1's dual role as codebook author and coder means the ensemble's closer agreement with the non-author HU2 is most plausibly evidence that the pipeline faithfully executes the written instrument rather than a sign of model inferiority; a third rater would settle this and is a low-cost addition. Finally, all four models are pinned to their routing on the date of the run, so AC1 values will not reproduce exactly on a future API version [38], which is why the per-call codes, not only the pipeline, are released. Beyond these, three constructive extensions follow from the findings: cross-register replication on citizen-deliberation transcripts, a structured argumentation-framework treatment of Counterarguments [8, 24], and a multi-agent consensus pipeline in which the models' rationales, not only their scores, are adjudicated [45].

6 Responsible Research

6.1 Ethics and research integrity

The executed study uses a two-coder, doubly-coded protocol to validate the LLM-assisted coding against independent human judgments, consistent with the emphasis on human validation in recent computational-content-analysis research [5, 19]. Two integrity limitations are owned: HU1 is both codebook author and coder — constrained but not removed by pre-registration, an independent non-author coder (HU2), and the release of all raw codes — and the second non-author coder that would separate faithful written-codebook application from a shared conservative reading could not be recruited. The configuration actually evaluated (a human lead coder for design and adjudication, with the LLM as second coder only) also preserves the methodological-apprenticeship value of double coding, which using an LLM as both coders would eliminate.

6.2 Data provenance and normative commitments

The data is exclusively public Hansard transcripts of UK House of Commons debate — a public record by long constitutional convention, raising no informed-consent issue — analysed at the level of speech acts by elected representatives in the course of public office; no third-party citizen named in debate is individually profiled. Codebook design embeds normative choices. Selecting Level of Justification, Respect for Groups, and Counterarguments reflects a pluralist theoretical position [6, 12, 18, 37, 53] and excludes

other legitimate conceptions of deliberative quality, particularly the Type II narrative-testimony and emotional-expression dimensions argued by Young [53] and Polletta and Lee [42]. The sub-codebook is a justified, tractable measurement instrument for one empirical context, not a complete theory of deliberative quality.

6.3 Use of generative AI

In line with the TU Delft guidelines on generative AI in BSc and MSc end projects, this section discloses how generative AI tools were used in the research and in writing this paper. Generative AI was used substantially, but as an assistant under continuous human direction and verification, not as a replacement for the author's own work. The LLMs evaluated as coders(Section 3.2) are the object of study, not a writing aid; their use is the research contribution itself and is documented in full in the methods and the reproducibility bundle. The use disclosed here is separate: the author's use of general-purpose AI assistants (principally large-language-model chat assistants and coding assistants) while conducting and writing the project.

Concretely, AI assistants were used for: *brainstorming and idea testing* — generating and stress-testing candidate research questions, framings, and counterarguments, which the author then accepted, rejected, or reshaped; *literature-search support* — surfacing candidate references and summarising known topics, with every cited source then located, read, and verified by the author against the original; *code assistance* — drafting, refactoring, and debugging the data-collection, LLM-pipeline, and analysis code, all of which the author reviewed, ran, and validated against expected behaviour; *data-analysis support* — drafting analysis and plotting scripts, with all reported numbers recomputed from the released data and checked by the author; *writing quality* — improving grammar, clarity, style, and readability of author-written drafts, and assisting with LaTeX formatting and table typesetting; and *presentation design* — structuring and visually refining the accompanying slides. No confidential or personal data was entered into any tool; the data is public Hansard text.

The author retains full intellectual and creative responsibility for the work. The core research idea, the conceptual research question, the three-dimension sub-codebook and its inclusion/exclusion warrants, the pre-registered analysis plan, the human coding, and the interpretation of the results are the author's own, and the author defined and enforced the methodological guardrails throughout — including, deliberately, the checks that guard against over-trusting LLM output, since the limits of LLM reliability are precisely this paper's subject. Generative AI can produce inaccurate, biased, or unverifiable content, including plausible but incorrect citations; accordingly, all AI-assisted text, code, and references were checked by the author, who is accountable for the accuracy, originality, and integrity of the final paper.

7 Conclusions

This paper investigated whether a four-model LLM ensemble can apply a theory-justified three-dimension DQI sub-codebook to parliamentary deliberation reliably enough to substitute for a trained human coder, and so scale deliberative-quality measurement beyond the reach of manual coding. To answer it, it paired a theory-justified sub-codebook (Level of Justification, Respect for Groups, Counterarguments) with a four-model LLM-as-judge ensemble at k = 5 self-consistency, and benchmarked the ensemble against two trained human coders on 200 UK House of Commons public-safety debate acts under Gwet's AC1 and a pre-registered paired test.

The answer to the question is: an LLM ensemble can substitute for a trained human coder on theory-grounded deliberative constructs that are well-specified and carry most

of their variance — here reason-giving and mutual recognition — but not on a construct whose written anchors fail to pin down a contested judgement, here engagement with counterarguments. In the case study, the ensemble substitutes as a second coder on every dimension; as a sole measure it substitutes on Level of Justification and Respect for Groups, while Counterarguments cannot yet be carried alone — and, in principle, the limit lies in the codebook design, not in the model, since the result analysis locates the fix in a revised anchor with worked examples. Establishing substantive downstream validity, transfer to citizen-deliberation registers, and independence from the single author-coder pair are the next steps toward LLM-scaled measurement of deliberative quality.

Acknowledgements

The author thanks Dr. William Brinkman and Michael Grauwde for their guidance throughout the project, and the trained collaborator who served as the second coder.

This project did not require ethical review. It analyses only the public UK House of Commons Hansard record, with no human participants and no personal or confidential data, so it falls outside the scope of the Human Research Ethics Committee process.

References

[1] Louis Abraham, Charles Arnal, and Antoine Marie. Prompt selection matters: Enhancing text annotations for social sciences with large language models. *Journal of Computational Social Science*, 8, 2025.

[2] André Bächtiger, Simon Niemeyer, Michael Neblo, Marco R. Steenbergen, and Jürg Steiner. Disentangling diversity in deliberative democracy: Competing theories, their blind spots and complementarities. *Journal of Political Philosophy*, 18(1):32–63, 2010.

[3] Maike Behrendt, Stefan Sylvius Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach, and Stefan Harmeling. AQuA: Combining experts' and non-experts' views to assess deliberation quality in online discussions using LLMs. In *Proc. DELITE @ LREC-COLING 2024*, pages 1–12, 2024.

[4] Simone Chambers. Deliberative democratic theory. *Annual Review of Political Science*, 6:307–326, 2003.

[5] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*, 2023.

[6] Joshua Cohen. Deliberation and democratic legitimacy. In Alan Hamlin and Philip Pettit, editors, *The Good Polity: Normative Analysis of the State*, pages 17–34. Basil Blackwell, 1989.

[7] John S. Dryzek. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford University Press, 2000.

[8] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[9] Zackary Okun Dunivin. Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170*, 2024.

[10] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[11] Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

[12] Nancy Fraser. Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text*, (25/26):56–80, 1990.

[13] Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. Deliberative abilities and influence in a transnational deliberative poll (EuroPolis). *British Journal of Political Science*, 48(4):1093–1118, 2018.

[14] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences (PNAS)*, 120(30):e2305016120, 2023.

[15] Anna Goddard and Alex Gillespie. Textual indicators of deliberative dialogue: A systematic review of methods for studying the quality of online dialogues. *Social Science Computer Review*, 41(6):2364–2385, 2023.

[16] Robert E. Goodin. *Innovating Democracy: Democratic Theory and Practice After the Deliberative Turn*. Oxford University Press, 2008.

[17] Kilem L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

[18] Jürgen Habermas. *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Beacon Press, 1984. Translated by Thomas McCarthy.

[19] Andrew Halterman and Katherine A. Keith. Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis*, 34(2):188–204, 2026.

[20] Michael Heseltine and Bernhard Clemm von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1), 2024.

[21] Hsiu-Fang Hsieh and Sarah E. Shannon. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288, 2005.

[22] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE, 4th edition, 2019.

[23] Ross Deans Kristensen-McLachlan, Miceal Canavan, Marton Kárdos, Mia Jacobsen, and Lene Aarøe. Are chatbots reliable text annotators? Sometimes. *PNAS Nexus*, 4(4):pgaf069, 2025.

[24] Sanjay Kumar, Jane Suiter, and Luca Longo. Advancing deliberative discourse measurement: The intersection with computational abstract argumentation in discourse quality evaluations. *Systems*, 13(3):204, 2025.

[25] Hao Lin and Yongjun Zhang. Navigating the risks of using large language models for text annotation in social science research. *Social Science Computer Review*, 44(3):403–427, 2026.

[26] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proc. EMNLP 2023*, pages 2511–2522, 2023.

[27] Christopher Lord and Dionysia Tamvaki. The politics of justification? applying the discourse quality index to the study of the european parliament. *European Political Science Review*, 5(1):27–54, 2013.

[28] Kathleen M. MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. Codebook development for team-based qualitative analysis. *Cultural Anthropology Methods*, 10(2):31–36, 1998.

[29] Rousiley C. M. Maia, Danila Cal, Janine Bargas, and Neylson J. B. Crepalde. Which types of reason-giving and storytelling are good for deliberation? assessing the discussion dynamics in legislative and citizen forums. *European Political Science Review*, 12(2):113–132, 2020.

[30] Jane Mansbridge. A minimalist definition of deliberation. In Patrick Heller and Vijayendra Rao, editors, *Deliberation and Development: Rethinking the Role of Voice and Collective Action in Unequal Societies*, pages 27–50. World Bank, 2015.

[31] Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F. Thompson, and Mark E. Warren. A systemic approach to deliberative democracy. In John Parkinson and Jane Mansbridge, editors, *Deliberative Systems: Deliberative Democracy at the Large Scale*, pages 1–26. Cambridge University Press, 2012.

[32] Jane Mansbridge, James Bohman, Simone Chambers, David Estlund, Andreas Føllesdal, Archon Fung, Cristina Lafont, Bernard Manin, and José Luis Martí. The place of self-interest and the role of power in deliberative democracy. *Journal of Political Philosophy*, 18(1):64–100, 2010.

[33] Sofie Marien, Ine Goovaerts, and Stephen Elstub. Deliberative qualities in televised election debates: The influence of the electoral system and populism. *West European Politics*, 43(6):1262–1284, 2020.

[34] Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. Are large language models reliable argument quality annotators? In *Proc. RATIO 2024, LNCS 14638*, pages 129–146, 2024.

[35] Patricia Mockler. Measuring the inclusiveness of deliberation: Structural inequality and the discourse quality index. *Comparative European Politics*, 20:53–72, 2022.

[36] Diana C. Mutz. *Hearing the Other Side: Deliberative versus Participatory Democracy*. Cambridge University Press, 2006.

[37] Simon Niemeyer, Francesco Veri, John S. Dryzek, and André Bächtiger. How deliberation happens. *American Political Science Review*, 118(1):345–362, 2024.

[38] Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*, 6:4–5, 2024.

[39] Nicholas Pangakis and Sam Wolken. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proc. Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 113–131. Association for Computational Linguistics, 2024.

[40] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*, 2023.

[41] Seraina Pedrini, André Bächtiger, and Marco R. Steenbergen. Deliberative inclusion of minorities: Patterns of reciprocity among linguistic groups in switzerland. *European Political Science Review*, 5(3):483–512, 2013.

[42] Francesca Polletta and John Lee. Is telling stories good for democracy? rhetoric in public deliberation after 9/11. *American Sociological Review*, 71(5):699–721, 2006.

[43] Michael V. Reiss. Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*, 2023.

[44] Maud Reveilhac. Comparing and mapping difference indices of debate quality on Twitter. *Methodological Innovations*, 16(2):234–249, 2023.

[45] Sebastian Simon, Sreecharan Sankaranarayanan, Elham Tajik, Conrad Borchers, Bahar Shahrokhian, Francesco Balzan, Sebastian Strauß, Sree Aurovindh Viswanathan, Amine Hatun Ataş, Mia Čarapina, Li Liang, and Berkan Celik. Comparing a human's and a multi-agent system's thematic analysis: Assessing qualitative coding consistency. In *Proc. AIED 2025, LNAI 15879*, pages 60–73, 2025.

[46] Marco R. Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1):21–48, 2003.

[47] Jürg Steiner, Maria Clara Jaramillo, Rousiley C. M. Maia, and Simona Mameli. *Deliberation Across Deeply Divided Societies: Transformative Moments*. Cambridge University Press, 2017.

[48] Petter Törnberg. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.

[49] Petter Törnberg. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195, 2025.

[50] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain-of-thought reasoning in language models. In *Proc. ICLR*, 2023.

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022.

[52] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. A comparison of cohen's kappa and gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13:61, 2013.

[53] Iris Marion Young. *Inclusion and Democracy*. Oxford University Press, 2000.

[54] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[55] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.

A LLM coding prompts and pipeline

This appendix contains the full prompts used at both stages of the G-Eval pipeline for each retained DQI dimension. Each dimension has a Stage A prompt (evaluation-step derivation, T = 0.0, cached) and a Stage B prompt (scoring, T = 0.7, k = 5 samples). All prompts are reproduced exactly as called via the OpenRouter API; placeholders in angle brackets are filled at call time. Figure A.1 shows the end-to-end pipeline referenced in Section 3.2.

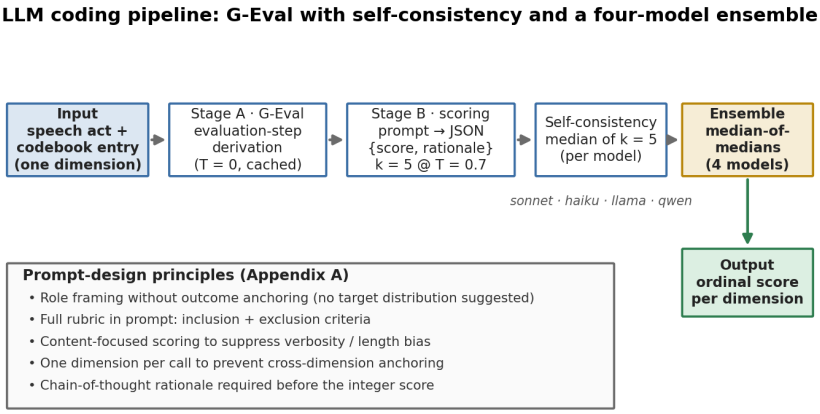

Figure A.1: The LLM coding pipeline. Each speech act is scored one dimension at a time with a two-stage G-Eval prompt, k = 5 self-consistency sampling per model, and a median-of-medians aggregation across four models.

A.1 System prompt (shared)

You are an expert discourse analyst applying the Discourse Quality Index to UK parliamentary debate transcripts. You score each speech act on a single ordinal dimension at a time, using only the codebook entry provided. Score the semantic content of the utterance. Do not score length, fluency, partisan affiliation, or speaker identity. When uncertain between two adjacent anchors, choose the lower anchor (conservative coding rule). Provide your rationale before the integer score. Output is single-line JSON only.

A.2 Stage A prompt: evaluation-step derivation

Given the following codebook entry for the dimension <DIMENSION_NAME>, generate a numbered list of 5 to 7 evaluation steps that a discourse analyst would apply, in order, to a single speech act to arrive at an ordinal score on this dimension. Each step should refer back to the codebook anchor language. Do not produce a score; produce only the evaluation steps. --- CODEBOOK ENTRY --- <CODEBOOK_ENTRY> --- END CODEBOOK ENTRY ---

A.3 Stage B prompt: scoring

You are scoring the following speech act on the dimension <DIMENSION_NAME> using the codebook entry and the evaluation steps below. --- CODEBOOK ENTRY --- <CODEBOOK_ENTRY> --- END CODEBOOK ENTRY --- --- EVALUATION STEPS --- <STAGE_A_OUTPUT> --- END EVALUATION STEPS --- --- SPEECH ACT --- <SPEECH_ACT> --- END SPEECH ACT --- Apply the evaluation steps in order. Write a brief rationale (2-4 sentences) citing the anchor language the speech act matches. Then output the final integer score. Output schema (single-line JSON, no Markdown fences): {"rationale": "<your rationale>", "score": <integer>} Constraints: - score must be in the dimension's ordinal range (LJ: 0-3; RG: 0-2; CA: 0-2). - rationale must precede score and must cite anchor language. - if uncertain between two adjacent anchors, choose the lower one.

A.4 Codebook entry example (Level of Justification)

Each codebook entry follows the MacQueen et al. [28] six-component structure (name, brief definition, full definition, when to use, when not to use, example). The Level of Justification entry is reproduced below; the Respect for Groups and Counterarguments entries are in the reproducibility bundle (codebook §2).

Name: Level of Justification (LJ). Scale: 0 (no justification), 1 (inferior), 2 (qualified), 3 (sophisticated). Brief definition: the extent to which the speaker provides reasons for their position. Full definition: a justification is any clause that gives a reason for the speaker's claim. A reason can be empirical, normative, or procedural. A bare assertion is LJ = 0; a single-clause weak reason is LJ = 1; a recognisable argument with one or more identifiable premises is LJ = 2; a multi-premise argument with explicit linkages is LJ = 3. When to use: every speech act on which the speaker advances a claim. When not to use: pure procedural utterances (point of order, division call), interjections under 30 words. Example: 'We must ban facial recognition because NIST has shown 35x higher false-match rates on darker-skinned women than on lighter-skinned men, and the Equality Act 2010 makes such disparate outcomes unlawful for public bodies.' -> LJ = 3.

B Supplementary analysis

B.1 Full dimension-selection justification

This section gives the full inclusion and exclusion reasons summarised in Section 3.1 and Figure 2.

Level of Justification (included). Reason-giving is the foundational claim of deliberative democracy. Habermas [18] requires that legitimate collective decisions be reached through reason-giving rather than mere assertion; Cohen [6] specifies free and reasoned agreement among equals; and Niemeyer et al. [37] operationalise the same idea in their Deliberative Reason Index. The empirical literature concurs: Pedrini et al. [41] find justification carries most of the variance in Swiss parliamentary deliberation, Steiner et al. [47] report it as the strongest predictor of deliberative transformative moments in post-conflict settings, and Gerber et al. [13] find rational justification the strongest opinion-change predictor in the EuroPolis poll. Behrendt et al. [3] report it among the harder DQI indicators to automate.

Respect for Groups (included). Young [53] argues that communication denigrating social groups constitutes communicative injustice; Fraser [12] identifies recognition as a prerequisite for participatory parity; and Mansbridge [30] and Chambers [4] treat mutual recognition as constitutive of deliberative legitimacy. In the public-safety AI context the dimension is particularly salient because predictive-policing, recidivism, and

biometric-surveillance systems disproportionately affect minority groups [10]. Operationally, Respect has the most concrete coding criteria of the six dimensions.

Counterarguments (included). Cohen [6] requires deliberators to remain open to revision in light of others' reasons, and Bächtiger et al. [2] identify counterargument engagement as the marker that most clearly separates deliberation from debate. Steiner et al. [47] include counterarguments in their reduced three-dimension DQI for citizen deliberation, a direct precedent for the present sub-codebook. Coding it requires high-level pragmatic inference, making it the most demanding LLM-evaluation challenge in the codebook.

Participation (excluded). Turn-taking is procedural and better evaluated at the system level [31, 35]. In Hansard, turn allocation is procedurally constrained by the Speaker, leaving the dimension near time-invariant.

Content of Justification (excluded). This dimension treats common-good appeals as the highest anchor — a ranking contested by Mansbridge et al. [32] and Young [53], destabilised across contexts by Maia et al. [29], and shown to be the weakest empirical performer by Reveilhac [44].

Constructive Politics (excluded). This dimension codes consensus-orientation. Goodin [16] decouples deliberative quality from constructive-politics outcomes, and Marien et al. [33] report direct floor effects on this dimension in adversarial multi-party television debates — the register Hansard occupies.

B.2 Per-coder label distributions

Table B.1: Per-coder per-dimension label distribution (n = 200 acts). The ENSEMBLE row is the median-of-medians; "—" marks anchors not on a dimension's scale.

Coder	Dim	0 — n (%)	1 — n (%)	2 — n (%)	3 — n (%)
HU1	LJ	3 (1.5)	37 (18.5)	141 (70.5)	19 (9.5)
HU1	RG	6 (3.0)	178 (89.0)	16 (8.0)	—
HU1	CA	101 (50.5)	4 (2.0)	95 (47.5)	—
HU2	LJ	21 (10.5)	67 (33.5)	111 (55.5)	1 (0.5)
HU2	RG	0 (0.0)	188 (94.0)	12 (6.0)	—
HU2	CA	186 (93.0)	12 (6.0)	2 (1.0)	—
sonnet	LJ	38 (19.0)	44 (22.0)	90 (45.0)	28 (14.0)
sonnet	RG	5 (2.5)	153 (76.5)	42 (21.0)	—
sonnet	CA	137 (68.5)	25 (12.5)	38 (19.0)	—
haiku	LJ	58 (29.0)	42 (21.0)	88 (44.0)	12 (6.0)
haiku	RG	5 (2.5)	166 (83.0)	29 (14.5)	—
haiku	CA	107 (53.5)	22 (11.0)	71 (35.5)	—
llama	LJ	26 (13.0)	18 (9.0)	156 (78.0)	0 (0.0)
llama	RG	8 (4.0)	152 (76.0)	40 (20.0)	—
llama	CA	109 (54.5)	50 (25.0)	41 (20.5)	—
qwen	LJ	34 (17.0)	21 (10.5)	128 (64.0)	17 (8.5)
qwen	RG	6 (3.0)	155 (77.5)	39 (19.5)	—
qwen	CA	122 (61.0)	61 (30.5)	17 (8.5)	—
ENSEMBLE	LJ	41 (20.5)	38 (19.0)	114 (57.0)	7 (3.5)
ENSEMBLE	RG	7 (3.5)	162 (81.0)	31 (15.5)	—
ENSEMBLE	CA	130 (65.0)	43 (21.5)	27 (13.5)	—

B.3 The kappa paradox

Figure B.1 plots the same coding data under Cohen's weighted kappa and Gwet's AC1, illustrating the artefact discussed in Section 3.3: under kappa every pair looks unreliable, while under AC1 the same pairs are substantial-to-almost-perfect. The collapse is sharpest on Respect for Groups, where about 90% of acts sit at a single anchor.

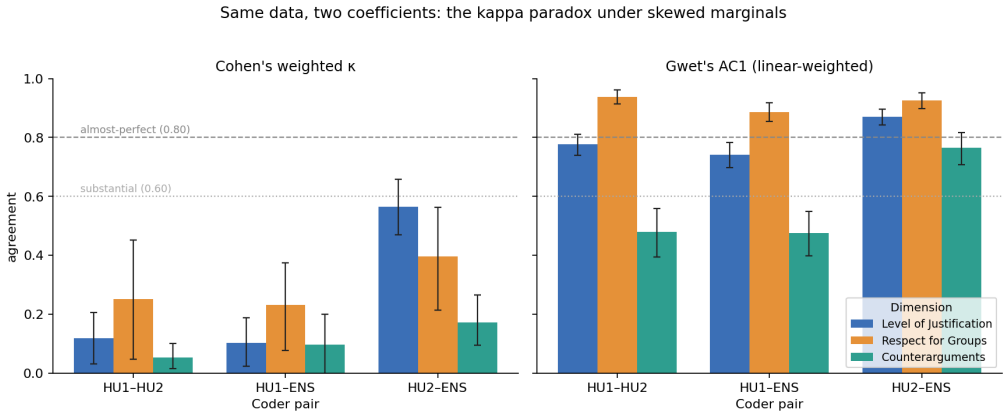

Figure B.1: The same coding data scored two ways. Under Cohen's weighted κ (left) every pair looks unreliable; under Gwet's AC1 (right) the same pairs are substantial-to-almost-perfect. Error bars are 95% bootstrap CIs.

B.4 Per-model and inter-model reliability

No single LLM dominates on every (dimension, human) cell: llama and qwen lead HU1 AC1 on Level of Justification, haiku leads HU2 AC1 on Respect for Groups, and qwen and sonnet lead HU2 AC1 on Counterarguments (Table B.2, Figure B.2). The inter-model comparison is reported because it is the diagnostic for the Counterarguments shortfall, not a redundant restatement of the per-model table: inter-model AC1 is uniformly substantial — every pair of active models exceeds 0.71 on every dimension and 0.92 on Respect for Groups (Table B.3) — so the four models share a more cohesive reading of the anchors than the two humans do on Counterarguments, which locates the binding constraint in the shared anchor language rather than in any one model. Ensembling delivers only modest gains over the best single model but is robust to model-specific bias in a way no single model is.

Table B.2: Per-LLM agreement with each human coder (n = 200; $AC1_{vv}$ with 95% bootstrap CIs). The gemini row is the dropped fifth model's supplementary partial sample (n = 22–24).

Model	Dim	vs. HU1 — AC1 (95% CI)	vs. HU2 — AC1 (95% CI)
sonnet	LJ	0.717 (0.675, 0.757)	0.821 (0.790, 0.851)
sonnet	RG	0.872 (0.836, 0.905)	0.900 (0.871, 0.929)
sonnet	CA	0.510 (0.432, 0.585)	0.744 (0.681, 0.802)
haiku	LJ	0.679 (0.638, 0.723)	0.806 (0.775, 0.838)
haiku	RG	0.893 (0.859, 0.923)	0.936 (0.909, 0.958)
haiku	CA	0.510 (0.432, 0.581)	0.581 (0.505, 0.652)
llama	LJ	0.811 (0.772, 0.847)	0.882 (0.855, 0.908)
llama	RG	0.853 (0.814, 0.887)	0.892 (0.859, 0.923)
llama	CA	0.468 (0.397, 0.536)	0.660 (0.596, 0.726)
qwen	LJ	0.771 (0.728, 0.811)	0.861 (0.834, 0.889)
qwen	RG	0.867 (0.828, 0.903)	0.906 (0.875, 0.934)
qwen	CA	0.460 (0.382, 0.531)	0.769 (0.724, 0.814)
gemini (supp.)	LJ	0.723 (0.596, 0.846)	0.825 (0.753, 0.912)

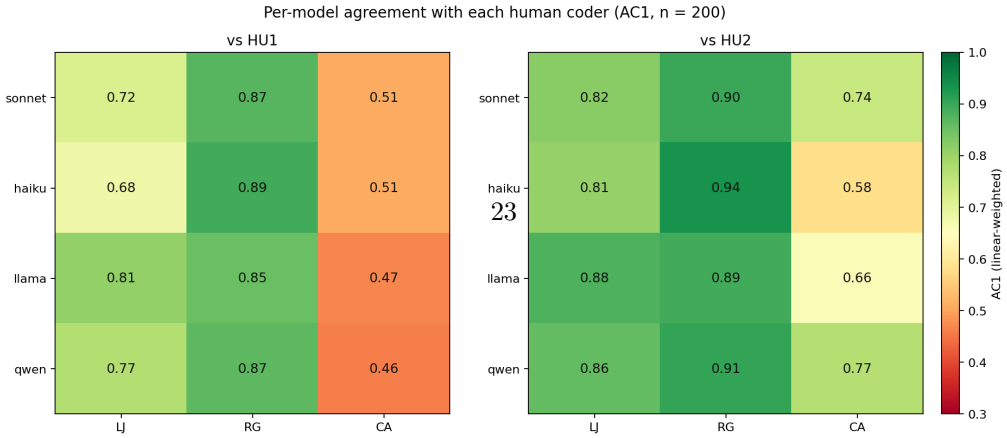

Table B.3: Inter-model reliability for representative pairs of active LLMs (n = 200; $AC1_{vv}$ with 95% bootstrap CIs); the full 18-row matrix is in the reproducibility bundle.

Model A	Model B	Dim	$AC1_{vv}$ (95% CI)
sonnet	haiku	LJ	0.850 (0.821, 0.881)
sonnet	haiku	RG	0.937 (0.911, 0.961)
sonnet	haiku	CA	0.756 (0.697, 0.818)
sonnet	llama	LJ	0.843 (0.814, 0.871)
sonnet	qwen	LJ	0.878 (0.851, 0.904)
sonnet	qwen	CA	0.767 (0.713, 0.814)
haiku	llama	CA	0.730 (0.677, 0.785)
haiku	qwen	RG	0.942 (0.917, 0.964)
llama	qwen	LJ	0.901 (0.874, 0.929)
llama	qwen	RG	0.953 (0.929, 0.972)
llama	qwen	CA	0.807 (0.762, 0.849)

B.5 Where the disagreement falls

The disagreement the AC1 matrix reports only in aggregate localises cleanly by ordinal distance. On Respect for Groups the mass sits on the diagonal (HU1–HU2 exact agreement 88%); on Level of Justification disagreement is adjacent, concentrated at the 1↔2 boundary (43 of 200 acts); on Counterarguments it is categorical, with HU1 coding CA = 2 where HU2 codes CA = 0 on 85 of 200 acts. Table B.4 decomposes each ensemble–human disagreement by ordinal distance.

Table B.4: Ordinal-distance decomposition of ensemble–human disagreement (n = 200). Exact, adjacent (±1), and gross (≥2 anchors apart) shares sum to 1 per row.

Pair	Dim	Exact	Adjacent (±1)	Gross (≥2)	Gross (% of disagr.)
HU1–ENS	LJ	0.485	0.360	0.155	30%
HU1–ENS	RG	0.800	0.190	0.010	5%
HU1–ENS	CA	0.450	0.215	0.335	61%
HU2–ENS	LJ	0.670	0.315	0.015	5%
HU2–ENS	RG	0.860	0.140	0.000	0%
HU2–ENS	CA	0.670	0.235	0.095	29%

B.6 Counterarguments failure-mode typology

The ensemble's divergences from HU1 on Counterarguments are not random noise but a small set of recurring rhetorical moves the codebook anchor does not adjudicate, read from the per-act rationales logged for every code (Table B.5).

Table B.5: Failure-mode typology of ensemble–HU1 divergence on Counterarguments, read from the per-act LLM rationales. Dir. is the direction relative to HU1 (under- or over-credit); n is the act count among the 110 Counterarguments (103 Level-of-Justification, for F4) ensemble–HU1 disagreements.

#	Failure mode	Divergence from HU1	Dir.	n
F1	Strict attribution	Scores a counter absent unless an opposing position is explicitly named	under	55

F2	Acknowledged, not engaged	Grants CA = 1 but withholds CA = 2 absent a restated rebuttal	under	25
F3	Concessive over-read	Reads the speaker's own "yes X, but Y" balancing as engaging a counter	over	28
F4	Thin-link over-grading (LJ)	Grades an implicit or rhetorical "because" chain as a qualified justification	over	14
F5	Anchor instability	One act draws three model codes; one model draws different codes across samples	both	106 / 21

Two acts localise the ambiguity. On a separation-centre answer that HU1 coded CA = 2, the four models returned three codes: haiku read "reasoned responses to the underlying objection" (CA = 2), sonnet "named only to set up the speaker's own reassurance" (CA = 1), and qwen "does not engage with any opposing position" (CA = 0) — splitting precisely over whether answering an unvoiced objection counts. On a probation answer, sonnet scored the same dismissive reference to a predecessor as CA = 1 ("a rhetorical foil") on one draw and CA = 0 ("falls under the exclusion criterion") on the next, on near-identical reasoning. The boundary is under-determined within a single coder, not only between coders — which is why the remedy is anchor language, not model capability.

B.7 Mirror substitution test

To confirm that the substitution conclusion does not hinge on anchoring the baseline on HU1, the mirror statistic $\delta' = \text{AC1}(\text{LLM ensemble, HU2}) - \text{AC1}(\text{HU1, HU2})$ re-runs the paired test against HU2 (Table B.6), with the same paired bootstrap (B = 1,000, seed 20260519) as Table 4. Anchored on HU2, the ensemble is statistically indistinguishable from the human baseline on Respect for Groups and significantly above it on Level of Justification and Counterarguments; the large Counterarguments value mainly reflects HU1's distinctive, permissive Counterarguments coding rather than super-human model performance. Because the ensemble agrees least with HU1, the HU1-anchored δ of Table 4 is the more conservative of the two anchorings.

Table B.6: Mirror paired $\delta' = \text{AC1}(\text{LLM ensemble, HU2}) - \text{AC1}(\text{HU1, HU2})$ per dimension, anchoring the baseline on HU2 rather than HU1.

Dim	**n**	**δ' AC1 (95% CI)**	**CI excludes 0**
LJ	200	+0.093 (+0.050, +0.134)	yes
RG	200	−0.012 (−0.042, +0.021)	no
CA	200	+0.286 (+0.192, +0.381)	yes

B.8 Variance and execution deviations

Within- and cross-model variance. Within-model self-consistency is high: across the 2,400 (act × dim × model) cells with k ≥ 4 samples the median standard deviation is 0.00 on every dimension, and most cells return an identical ordinal label on all five samples. The dominant uncertainty signal is instead cross-model disagreement (Figure B.3): per-act cross-model vote entropy has median 0.81 bits on Level of Justification, 0.77 on Counterarguments, and 0.00 on Respect for Groups. A post-hoc check (outside the pre-registered plan) found that cross-model entropy does not flag the acts humans find hard — it predicts HU1–HU2 disagreement at AUROC near chance on Level of Justification and Respect for Groups — so it is at most a weak, dimension-specific confidence signal, not a general triage mechanism.

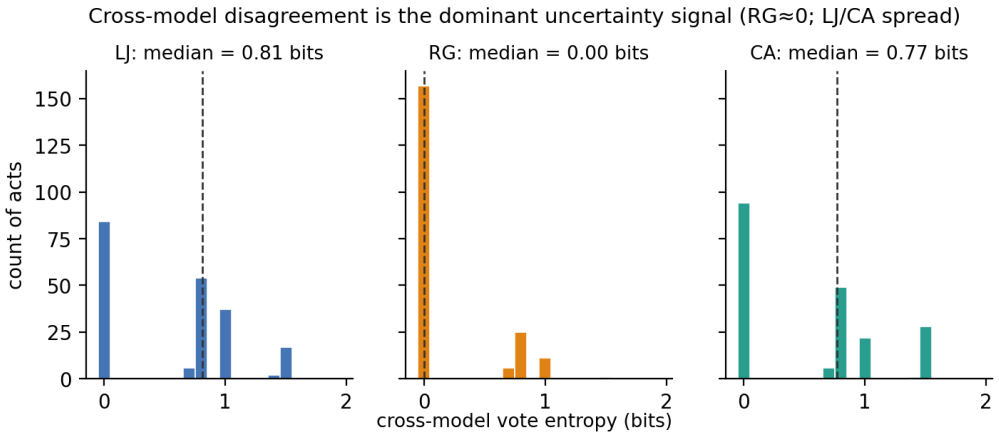

Figure B.3: Cross-model vote entropy per act. Disagreement between the four models, not within-model sampling noise, is the dominant source of uncertainty: entropy is near 0 on Respect for Groups and spread on Level of Justification and Counterarguments.

Deviations from the pre-registered plan. Four deviations from the pre-registration (v1, time-stamped 2026-05-19) are recorded; all are reported here rather than folded silently into the design. (i) *Headline metric.* The pre-registration named Cohen's linear-weighted kappa as the primary statistic and Gwet's AC1 as a secondary high-prevalence-paradox guard. Because the paradox in fact materialised on these skewed marginals (Appendix B.3), AC1 is promoted to the reported headline and kappa is retained alongside it for back-compatibility; the substitutability decision is reported on AC1. (ii) *Substitutability operationalisation.* The pre-registration set the substitutability bound as a ±0.10 band on the agreement coefficient and operationalised the decision as interval overlap with the human baseline; we report the same ±0.10 bound in the stricter and more standard paired-δ equivalence form (the CI for the paired difference lying inside the band), which is the form used throughout Section 4. (iii) *Decoding and ensemble.* The pre-registration described single-pass coding at temperature 0.0 with a primary model plus one comparator (and up to a five-model pool). In execution the pipeline uses a four-model ensemble (claude-sonnet-4.6, claude-haiku-4.5, llama-3.3-70b, qwen-2.5-72b) with $k = 5$ self-consistency sampling at temperature 0.7, the cached evaluation-step derivation at T = 0.0, OpenRouter seed = 20260519 + sample_k, max_tokens = 400, and three retries per call; gemini-2.5-pro was trialled as a fifth model but dropped after consuming disproportionate budget on 22–24 of 200 acts (its partial samples are a supplementary comparator, Table B.2), and the originally named gpt-4o-mini comparator was not used. (iv) *Tie-break.* Aggregation applies the codebook's conservative tie-break (lower median

when adjacent anchors tie) at both the within-model and across-model layers. Total completed calls: 12,274 (11,999 on the four active models plus 275 supplementary gemini samples), each logged with timestamp and content hash. All 2,400 (act × dim × model) cells have at least $k = 4$ samples; 2,399 have the full $k = 5$. The pre-registered sampling seed, the 200-act calibration set, the three retained dimensions, the codebook-stability trigger, and the human double-coding protocol are unchanged from the plan.

B.9 Reasoning-faithfulness audit

Each scoring call returns a written rationale alongside its integer score (Appendix A). The faithfulness question is whether the rationale concludes the anchor the call actually emits — the property a logged rationale is supposed to guarantee, and the one Lin and Zhang [25] verify by hand and Törnberg [49] cautions may be post-hoc. A rule-based extractor reads the concluded anchor from each rationale: it accepts an "<DIM> = n" mention only when an assertion cue (*satisfying, meets, fits, aligns with, scores*) immediately precedes it and no negation or hypothetical cue (*not, below, would require, to reach*) is adjacent, taking the last such assertion as the conclusion. This is deliberately high-precision and low-coverage: it reads a concluded anchor on 4,118 of the 11,999 ensemble calls (34%). On those calls the extractor's anchor matches the emitted score in 88.5% of cases (Table B.7). A by-hand audit of a 50-call stratified random sample (seed 20260519) found that this 88.5% understates the true rate: all eight automated "mismatches" in the sample were extraction errors in which the rationale stated a contrastively *rejected* adjacent anchor ("matching the language for LJ = 2, but not meeting the criteria for LJ = 3") that the extractor mistook for the conclusion, and none was a genuine score–rationale contradiction. The audit thus found 50/50 calls faithful, putting a Wilson 95% upper bound of roughly 7% on the genuine inconsistency rate. Per-model differences in the automated rate track phrasing style — models that narrate rejected anchors before concluding trip the extractor more often — and are not interpreted as capability differences. The headline reading is that the ensemble's scores are consistent with their stated reasoning; faithfulness in this sense is necessary, not sufficient, for the rationale to be the score's actual cause.

Table B.7: Automated reasoning-faithfulness rate per dimension: fraction of calls (of those with a machine-readable concluded anchor) whose emitted score equals the rationale's concluded anchor. Wilson 95% CIs. A 50-call manual audit found the residual mismatches to be extraction artefacts, so these are conservative floors.

Dimension	Calls read	Coverage	Faithful (95% CI)
Level of Justification	1,636	0.41	**0.913** (0.898, 0.925)
Respect for Groups	1,150	0.29	**0.839** (0.817, 0.859)
Counterarguments	1,332	0.33	**0.891** (0.873, 0.907)
All dimensions	4,118	0.34	**0.885** (0.875, 0.895)

B.10 Contamination / memorisation check

Because the Hansard corpus is public, the pre-2025 acts are plausibly present in the four models' pre-training data, raising the standard contamination concern [1]: that agreement reflects recall of seen text rather than application of the rubric. Two features of the design bound this. First, the DQI scores are not in any pre-training corpus — they were generated by HU1 and HU2 for this study and never published — so no label can have been memorised; only the debate text is a candidate. Second, the corpus spans 2020–2025, with 81 of the 200 acts drawn from 2025, at or beyond the models' training windows and therefore unseeable. Under a memorisation account, ensemble–human agreement would fall on these recent, unseen acts; it does not (Figure B.4). The Spearman correlation

between debate year and per-act ensemble–HU1 exact agreement is $\rho = 0.03$ (p = 0.70); mean exact agreement is 0.577 on the 119 acts from 2020–2024 and 0.580 on the 81 acts from 2025 (paired-resample difference −0.003, 95% CI −0.078 to +0.072). Computed as weighted AC1 against HU1 and split at the 2025 boundary, no dimension degrades on the unseen acts (LJ 0.734 → 0.752; RG 0.892 → 0.879; CA 0.457 → 0.504). The flat recency profile is the signature of rubric application, not text recall.

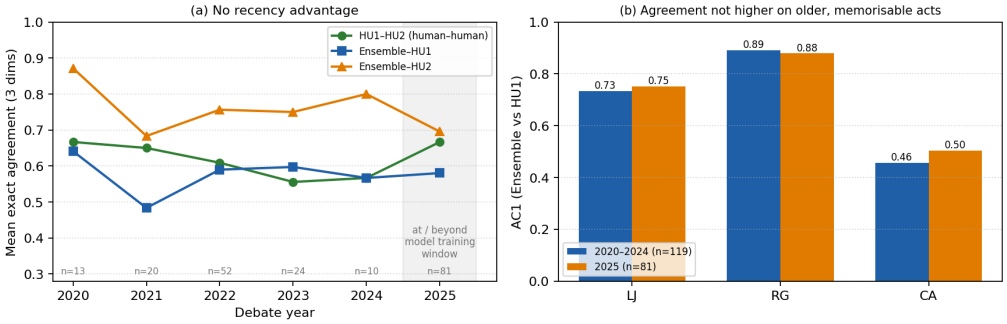

Figure B.4: Contamination check. (a) Mean exact agreement (across the three dimensions) by debate year for the human–human, ensemble–HU1, and ensemble–HU2 pairs; the shaded band marks acts at or beyond the models' training windows. Agreement shows no recency advantage. (b) Weighted AC1 of the ensemble against HU1 per dimension, split into 2020–2024 (seen) and 2025 (unseen) acts; agreement does not fall on the unseen acts.

B.11 Robustness of the substitution result

The headline substitution δ (Table 4) is re-estimated under perturbations of the three pipeline choices the executed design can vary without new model calls: stochastic decoding, model composition, and the aggregation rule. Within-model self-consistency at $k = 5$, $T = 0.7$ is high — the five samples are unanimous on 69.1% (LJ), 85.1% (RG), and 68.5% (CA) of the 800 (act × model) cells per dimension, with mean modal-vote shares of 0.90, 0.96, and 0.91 and a median within-model standard deviation of zero — so stochastic-decoding sensitivity is small. Dropping each model in turn (Table B.8) and replacing the median-of-medians rule with a mean-then-round, a majority mode, or a pooled median over all twenty samples (Table B.9) leaves all 24 paired δ point estimates inside the pre-registered ±0.10 band; Counterarguments stays statistically indistinguishable from the human baseline under every variant, and the Level-of-Justification and Respect-for-Groups deltas keep the same sign and rough magnitude as the headline (Figure B.5). No single provider and no single aggregation choice drives the result. The untested axis is prompt *wording*: re-running under paraphrased prompts was outside the API budget, and Abraham et al. [1] show prompt phrasing can move accuracy materially, so prompt-paraphrase robustness is the principal open question rather than a settled one.

Table B.8: Leave-one-model-out robustness. Paired δ = AC1(ensemble, HU1) − AC1(HU1, HU2) with each model removed from the four-model ensemble; paired bootstrap 95% CIs, seed 20260519.

Model dropped	Dim	δ AC1 (95% CI)	In band
sonnet	LJ	−0.002 (−0.034, +0.032)	yes
sonnet	RG	−0.059 (−0.089, −0.031)	yes

sonnet	CA	−0.009 (−0.083, +0.059)	yes
haiku	LJ	−0.001 (−0.034, +0.035)	yes
haiku	RG	−0.076 (−0.108, −0.045)	yes
haiku	CA	−0.001 (−0.069, +0.068)	yes
llama	LJ	−0.049 (−0.085, −0.015)	yes
llama	RG	−0.053 (−0.081, −0.026)	yes
llama	CA	+0.025 (−0.048, +0.096)	yes
qwen	LJ	−0.023 (−0.058, +0.011)	yes
qwen	RG	−0.059 (−0.089, −0.031)	yes
qwen	CA	+0.007 (−0.073, +0.086)	yes

Table B.9: Aggregation-rule robustness. Paired δ under the headline median-of-medians rule and three alternatives; paired bootstrap 95% CIs. Counterarguments is indistinguishable from the human baseline under every rule.

Aggregation rule	**Dim**	**δ AC1 (95% CI)**	**In band**
Median-of-medians (headline)	LJ	−0.036 (−0.068, −0.003)	yes
Median-of-medians (headline)	RG	−0.051 (−0.078, −0.023)	yes
Median-of-medians (headline)	CA	−0.002 (−0.065, +0.059)	yes
Mean-then-round	LJ	−0.004 (−0.039, +0.029)	yes
Mean-then-round	RG	−0.076 (−0.112, −0.043)	yes
Mean-then-round	CA	+0.011 (−0.066, +0.093)	yes
Majority mode	LJ	−0.030 (−0.066, +0.001)	yes
Majority mode	RG	−0.051 (−0.077, −0.024)	yes
Majority mode	CA	+0.002 (−0.070, +0.070)	yes
Pooled median (20 samples)	LJ	−0.017 (−0.051, +0.014)	yes
Pooled median (20 samples)	RG	−0.056 (−0.087, −0.029)	yes
Pooled median (20 samples)	CA	+0.010 (−0.058, +0.082)	yes

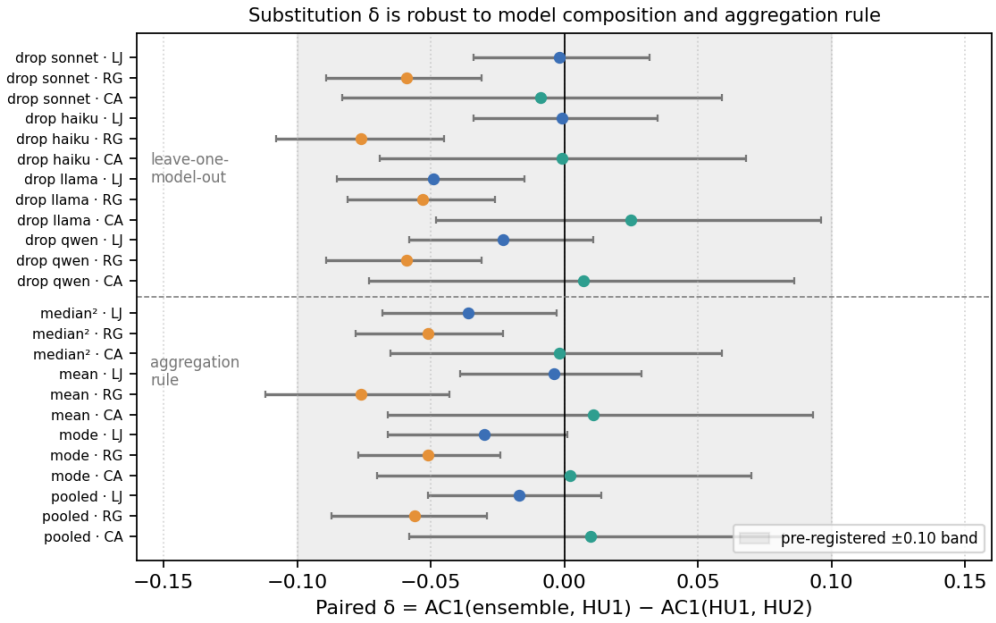

Figure B.5: Robustness of the substitution $\delta = \text{AC1}(\text{ensemble, HU1}) - \text{AC1}(\text{HU1, HU2})$. Each point is the paired δ under a leave-one-model-out ensemble (top) or an alternative aggregation rule (bottom), coloured by dimension, with paired-bootstrap 95% CIs. All 24 point estimates fall inside the pre-registered ± 0.10 band (shaded).