Deliberative, Not Merely Informed: Argument-Grounded Polling from Mined Deliberation

Lucas Anastasiou[1], Enrico Daga[1] and Anna De Liddo[1]

[1]*Knowledge Media Institute, The Open University, Milton Keynes, UK*

Abstract

Deliberative Polling measures how informed deliberation changes public opinion; however, its cost - representative samples, moderated face-to-face sessions - confines it to rare flagship events. In this work, we ask whether the argumentative content of small deliberative events can be extracted and structured well enough to be re-served to a wider audience, as the basis for asynchronous, *argument-grounded* polling. Working from six deliberative focus groups on urban policy held in Lisbon (Portuguese language, 35 participants), we develop an LLM-based pipeline that mines an argument graph of claims, premises, stances and speakers, serialised in the Argument Interchange Format (AIF) and with entity mentions linked to DBpedia and Wikidata. Lacking a human-coded gold standard, we mine each event independently with five LLM agents from distinct model families, report inter-agent agreement as a reliability signal, and construct a majority-vote consensus graph. From the graph we then generate narrative digests with every sentence traceable to source utterances, and serve them in a bilingual polling pilot ($N = 112$, powered for estimation) against a transcript-direct summary and a no-digest control. Overall, serving mined argumentation improves reader comprehension: compared to reading bare proposal descriptions, digest readers recognized actual arguments more accurately (0.63 vs. 0.47; $\delta = +0.41$) and endorsed fake arguments far less often (0.03 vs. 0.18; $\delta = +0.37$). However, graph grounding did not outperform direct LLM summarization at the reader level: results were flat in Portuguese and reversed in English due to an unintended length and balance mismatch in the English stimulus. On the generation side, graph grounding uniquely guarantees sentence-level verifiability that unsourced summaries cannot provide. We conclude by defining the core guarantees needed for a poll to count as *deliberative* rather than *merely informed*, showing that this distinction depends on serving argumentation itself rather than the specific text generation mechanism.

Keywords

Argument mining, Deliberative polling, Deliberative democracy, Civic technology, Knowledge graphs, Grounded generation, Data-driven storytelling

CEUR-WS.org/Vol-4313/dds-paper1.pdf

1. Introduction

Deliberative Polling [1, 2] is among the best-evidenced instruments in democratic innovation: a representative sample of citizens is convened, briefed with balanced materials, and moderated through structured discussion, and the movement between a pre- and a post-deliberation survey is read as the effect of *informed* opinion formation. Nonetheless, its evidential strength is also its constraint. Representative recruitment, trained moderation and multi-day sessions make each event expensive and rare, so the deliberative signal reaches only the room that produced it - a transmission bottleneck documented across the wider family of deliberative mini-publics [3, 4].

Recent work asks whether language models can lower the cost of the *deliberative act* itself, for instance by mediating discussion toward common ground [5]. We ask instead a complementary question on the *output* side: once a small deliberative event has happened, can its argumentative content be extracted and structured faithfully enough to be re-served to a wider audience, so that polling on that material approximates a Deliberative Poll at a fraction of the cost?

We argue that the load-bearing requirement of such a design is twofold: *argumentative exposure* to inform readers, and *structural grounding* to make the process auditable. While serving mined argumentation makes readers *deliberative* (accurate and resistant to false positions, rather than merely

Data-Driven Storytelling: Bridging Knowledge Graphs, GenAI, and Narrative (DDS2026), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

lucas.anastasiou@open.ac.uk (L. Anastasiou); enrico.daga@open.ac.uk (E. Daga); anna.deliddo@open.ac.uk (A. De Liddo)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

informed about bare proposals), the argument graph itself supplies system-level legitimacy through sentence-level provenance, structural balance enforcement, and cross-lingual entity anchoring.

We prototype this end-to-end on the Lisbon Deliberative Focus Groups (DFGs) of the MultiPoD project [1]: six moderated events on urban policy, in Portuguese. Our contributions are (i) a reusable pipeline that mines an AIF argument graph from colloquial spoken Portuguese and links its entities to DBpedia and Wikidata; (ii) a *gold-free* reliability methodology that mines with five LLM agents from distinct model families and constructs a majority-vote consensus graph; (iii) balanced, fully traceable narrative digests, with an ablation testing of graph grounding over direct summarisation; and (iv) an argument-grounded polling protocol, the guarantees a pipeline must meet for the resulting poll to count as deliberative, and a bilingual pilot We release the schema, mining and entity linking pipeline, narrative digests, survey instrument and response data[2].

Two results shape the paper. On the one hand, serving mined argumentation *does* make readers more accurate about a deliberation they never attended. On the other hand, the mechanism we hypothesised would add an extra accuracy - grounding generation in the argument graph rather than summarising the transcript - did not separate the two digest conditions at reader level.We report this negative result, diagnose the stimulus confound behind its English reversal, and demonstrate that while serving arguments drives reader-side accuracy, graph grounding earns its place on the generation side by guaranteeing auditability and balance enforcement.

2. Background and Related Work

Our proposal sits at the meeting point of several literatures that rarely cite one another. Deliberative Polling [1, 2] contributes the pre/post attitude-change design we borrow and the normative standard - balanced, informed exposure to the arguments - that our faithfulness requirement operationalises computationally. Its central unresolved question, however, is *transmission*, as mini-publics influence wider politics only through fragile uptake channels [3], a bottleneck that the OECD finds replicated across nearly three hundred institutional processes [4]. Technical responses have so far targeted the deliberative act itself, e.g. an LLM mediator that helps groups find common ground [5], or the Democratic Reflection interface, deployed in the events we study, which captures time-stamped audience reactions during deliberation [6], complementing digital deliberation platforms and sensemaking tools designed to scaffold structured discussion and reduce group bias [7]. We operate instead on the deliberation's *record*, making transmission itself the object of engineering and measurement. Turning that record into structure is the task of argument mining, i.e. the Turning that record into structure is the task of argument mining, i.e. the extraction of claims, premises and support/attack relations from text, for which spontaneous spoken dialogue remains the hardest register [8]. Recent work in deliberative technology has explored argument mining in structured, decentralised discussion platforms to support democratic deliberation [9]. We serialise our results in the Argument Interchange Format [10] standard and our graph is meant to be reused - extended with speaker, timing, stance and entity attributes. Where classical argument mining evaluates against human gold data, recent evidence that LLMs are capable text annotators [11] raises the question of what quality even means when the annotators are themselves models. We answer with the classical apparatus of inter-annotator reliability - that is, chance-corrected agreement across independent annotators [12, 13] - applied to five deliberately heterogeneous model families, and we hold the resulting claim to be one of *reliability*, and not of validity (Section 5).

Our digests are data-driven stories in the sense of the literature this workshop convenes, which has grown from narrative visualisation [14, 15] into a programme on generating narratives from structured data [16, 17, 18]. Knowledge graphs increasingly *steer* such generation, e.g. as controllable backbones for long-form stories [19], as knowledge scaffolds for data narration [20], and as ontological foundations for narrative itself [21]; moreover, surveys of LLM story generation identify structured planning as the

[1] https://multipod-project.eu/

[2] Replication package, mining prompts, consensus graphs, digests, and survey instrument: https://github.com/anastluc/argument-grounded-deliberation-polling

main lever on coherence [22, 23, 24]. The converse direction, i.e. LLMs constructing knowledge graphs from text including argumentative and scholarly discourse [25, 26], is exactly our extraction step. What sets our work apart, however, is that readers use these narrative digests to form *political* judgments. Every sentence must therefore trace back directly to what a participant said, making faithfulness an essential legitimacy requirement rather than an optional quality metric.

That condition connects us, in turn, to the faithfulness literature. Abstractive summarisation is known to hallucinate content unsupported by its source [27], and a decade of evaluation work has produced entailment-, QA- and LLM-judge-based metrics of factual consistency [28] alongside training-time mitigations [29]; in this line, knowledge graphs enter as *referees* that localise hallucinations [30] or supply gold references for multilingual evaluation [31]. Following recent multilingual data-to-text research, we prioritize *attributability*-ensuring every statement carries a direct, checkable pointer to its source evidence-over post-hoc hallucination detection [32]. Accordingly, our digest sentences cite the utterances they rest on, and our evaluation separates what can be verified from what merely cannot be checked. Finally, because our judges are themselves LLMs, we follow the known cautions on LLM-as-judge bias [33, 34] and use a judge from a different model family than the generator.

Lastly, anchoring mentions to knowledge-base identities is mature for prominent entities - DBpedia Spotlight [35] over DBpedia [36], and entity search over Wikidata [37] - but degrades on the long tail [38, 39]. Civic deliberation, however, is a classic example of a long-tail domain, as neighbourhoods, municipal programmes and colloquially named proposals dominate the discourse and are precisely what knowledge bases lack. Our two-base linking and the resulting gap-fill worklist (Section 5.3) therefore quantify this gap on deliberative speech.

3. Why Ground Polling in an Argument Graph?

The promise of argument-grounded polling is that the expensive part of a Deliberative Poll - a real, moderated exchange of reasons among citizens - happens only once, while the poll built on its arguments reaches arbitrarily many asynchronous respondents at marginal cost. Early explorations into automated reporting for online discussions demonstrated that structured summaries can assist citizens in navigating and making sense of complex deliberations [40]. For that transfer to inherit any deliberative legitimacy, the served material must preserve what was argued, in what proportion, and which positions had real constituencies. it by construction, and it does so in four ways: (i) stance-typed, proposal-clustered claims make proportional coverage a constraint the generator can be *held to*; (ii) source-utterance identifiers on every node make each digest sentence traceable to what was said; (iii) a discrete representation can be mined independently by several annotators and consensus-filtered, which free prose cannot; and (iv) knowledge-base anchors turn the discussion's referents into language-independent identities, making the record queryable and aggregable across events.

From this rationale we derived three pre-stated hypotheses: **H1** (exposure), that readers of either digest shift attitudes more than readers of bare proposal descriptions; **H2** (grounding, accuracy), that graph-digest readers recognise what was argued more accurately than readers of a transcript-direct summary; and **H3** (grounding, fabrication), that graph-digest readers less often endorse positions no participant held. In effect, H2 and H3 test whether the graph's balance and provenance constraints yield a *reader-visible accuracy premium* beyond the system-level auditability they inherently provide. As we show below, the pilot supports H1 (serving arguments vs. proposals) but neither H2 nor H3, locating the graph's primary value in system accountability and balance validation rather than reader-side differentiation (Section 8).

4. The Lisbon Deliberative Focus Groups

The corpus derives from the first citizen pilots of MultiPoD, an EU Horizon Europe project developing digital and hybrid tools for inclusive, multilingual political deliberation. Between November and December 2024 the project's Lisbon pilot team convened six Deliberative Focus Groups (DFG1–DFG6)

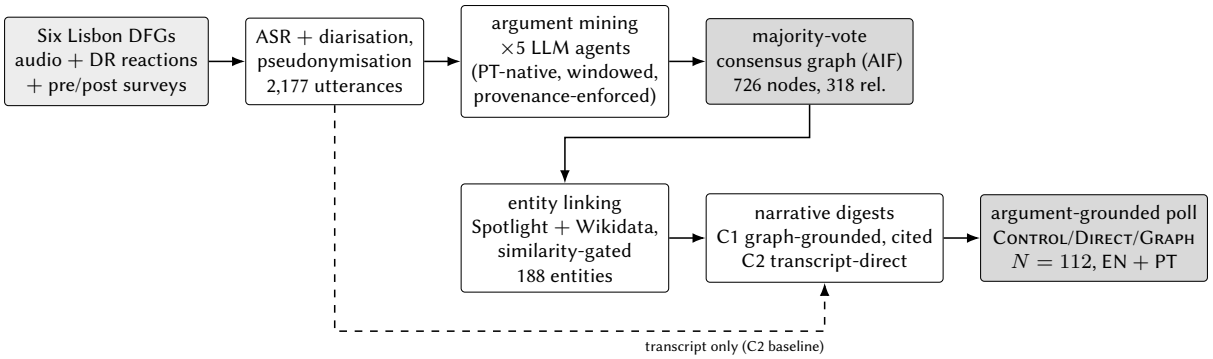


Figure 1: The pipeline. Five LLM agents from distinct model families independently mine each event; only structure at least three agree on enters the consensus graph, which is entity-linked and drives the graph-grounded digest (C1). The transcript-direct digest (C2), generated without the graph, is the ablation baseline; both are served, against a no-digest control, in the polling pilot. Each stage supplies one of the protocol's guarantees: consensus → argument fidelity, linking → entity fidelity, grounded generation → balance and provenance.

- small moderated assemblies, run in Portuguese, with 35 citizen participants in total across the six sessions. The deliberations addressed live questions of Lisbon urban policy, namely: what to do with the site of the Humberto Delgado airport if it moves to Alcochete, how to manage tourism pressure, public and affordable housing, and the provision of green space.

Each session followed the same protocol. After welcome and informed consent, participants recorded their baseline positions in a short attitude survey (support for each proposal on a 3-point scale) and received short, neutral written descriptions of the proposals - deliberately argument-free, and the same framing later reused for the pilot's control condition. A trained facilitator then introduced each proposal in turn and moderated an open discussion around it, supported by a second facilitator, with an observer taking field notes. Moderation aimed to draw out reasons rather than tallies, and these moderator introductions later serve as topic anchors for the mining pipeline. After the final discussion block participants repeated the attitude survey, yielding 591 pre/post responses over the six events. Throughout, participants could also react to the unfolding discussion on *Democratic Reflection*, a card-based interface originally developed for televised election debates [6], by tapping time-stamped evaluative cards; this deployment produced 3,375 card responses from the 11 participants who used it. Sessions were audio-recorded on two channels.

In this paper, in-room behavioural signals serve primarily as corpus context: while the mined graph includes per-utterance timestamps to allow future alignment between click responses and argument nodes, that alignment is outside our current dataset release, and we report no in-room click-survey statistics here.

Participation was covered by the project's informed-consent and data-protection procedures. All material used here derives from anonymised transcripts, i.e. speaker names and places of residence spoken aloud are removed and speakers are pseudonymised.

5. From Transcripts to an Entity-Linked Argument Graph

Figure 1 gives an overview of the pipeline. From the DFGs we derive the argument mining and entity linking pipeline, the generated narrative digests, and the survey response data - with code, digests, and survey data published in the project repository under CC BY 4.0 (raw transcripts and intermediate argument graphs are omitted to protect privacy). The prompts driving each LLM stage are reproduced in Appendix A.

5.1. Transcripts

Audio for all six events was transcribed with automatic speech recognition and speaker diarisation (AssemblyAI, language `pt`). For each event we select the more fully diarised of the two recording channels, concatenate its parts, and normalise every utterance to a common schema `{event_id, utterance_id, speaker_label, t_start, t_end, text}`. Speakers are pseudonymised per event and spoken names and locations stripped. The normalised corpus contains 2,177 utterances (Table 2). Diarisation errors on overlapping colloquial speech are accepted and noted as a limitation; we do not attempt a speaker-to-survey-respondent mapping, which the design does not require.

5.2. Multi-agent argument mining and the consensus graph

We serialise arguments in AIF [10]: I-nodes carry propositions (claims and premises), RA- and CA-nodes support and attack. We extend I-nodes with speaker, start/end time, a stance toward the proposal under discussion (pro/con/neutral), and `mentionsEntity` links to `mentionsEntity` links to knowledge-graph URIs; proposals are modelled as first-class nodes (that is, independent, addressable vertices in the AIF graph possessing unique URIs and metadata rather than transient edge labels, enabling relations, queries, and stance mappings to target them directly). The doubles as the schema artifact in the repository.

Extraction runs over sliding windows of $W = 15$ consecutive utterances with a 5-utterance overlap (stride $S = 10$), ensuring that argumentative moves spanning a boundary are seen whole at least once. This window-level extraction trades granularity for robustness, a deliberate choice on colloquial speech where utterance-level classification is brittle. Within each window the model segments and classifies units as claim/premise/other, attaches a stance and target proposal (the moderator's proposal introductions serve as topic anchors), and links support/attack relations within the topical segment (Appendix A). Prompting is Portuguese-native, and we never translate the corpus, so as to avoid compounding translation error with extraction error. Furthermore, provenance is enforced mechanically: any extracted unit that does not cite a source-utterance identifier present in its window is discarded, to guard against fabricated evidence at the extraction stage. Finally, overlap duplicates are merged, and proposal mentions are clustered into proposal nodes by accent-insensitive string matching (i.e. light deduplication, and not semantic merging).

No human-coded gold standard aligned with our schema exists for this corpus, and coding one at scale was out of scope. Therefore, rather than validate against a single transcript, we treat extraction as an inter-annotator reliability problem. Five LLM agents from distinct vendors - `openai/gpt-5.6-luna`, `qwen/qwen3.7-flash`, `google/gemini-3.1-pro-preview`, `anthropic/claude-opus-4.8` and `z-ai/glm-5.2`, a deliberate mix of proprietary and open-weight families - each mine every event under identical prompts, schema, and temperature ($T = 0.0$). Family diversity is the design choice doing the work here, as annotators sharing a training lineage would err in correlated ways and inflate agreement. We report chance-corrected agreement - Krippendorff's α and Fleiss' κ [12, 13] - separately per decision level, because the levels differ in difficulty (Table 1). We deduce that agreement is substantial for We deduce that agreement is substantial for argument detection ($\alpha = 0.70$), moderate for unit type ($\alpha = 0.62$) and stance ($\alpha = 0.56$), and near chance for support/attack relations ($\alpha = 0.08$), which is consistent with relation extraction being the hardest, least reproducible step on colloquial spoken Portuguese. We therefore treat detection and typing as reliable, relations as preliminary, and exclude low-agreement structure by consensus voting.

Error analysis of relation extraction failure modes. The low inter-agent agreement on relation extraction ($\alpha = 0.08$) reflects three characteristic failure modes of ensemble LLMs on spontaneous conversational speech: (i) *Long-distance conversational dependencies*: participants in focus groups frequently challenge or support remarks made several turns earlier, often beyond the local sliding extraction window ($W = 15$); agents consequently bind relations to divergent local antecedents or miss the distal anchor entirely. (ii) *Concessive and ambiguous polarity*: colloquial European Portuguese

Table 1
Inter-agent reliability across the five agents (all events). α: Krippendorff's alpha; κ: Fleiss' kappa; raw: mean pairwise raw agreement.

Decision level	α	κ	raw
Argument detection	0.70	0.70	0.87
Unit type (claim/prem.)	0.62	0.62	0.81
Stance (pro/con/neut.)	0.56	0.56	0.77
Support/attack relations	0.08	0.08	0.54

frequently blends concession with disagreement (e.g., *"compreendo perfeitamente, mas..."* / *"I completely understand, but..."*); different model families systematically diverge on whether to classify such moves as supportive acknowledgment or adversarial rebuttal. (iii) *Implicit warrants (enthymemes)*: everyday civic debate routinely omits explicit deductive steps, forcing models to impute unstated inferential warrants, yielding heterogeneous relation predictions across prompts. Crucially, our downstream pipeline is insulated from this relation noise: digest generation clusters propositions primarily by target proposal and node-level stance rather than deeply nested relational trees, preventing relation instability from corrupting narrative synthesis.

From the five agent graphs we then build a majority-vote *consensus* graph: a node is retained when at least three of five agents extract an argument covering the same source utterance(s), with unit type and stance decided by vote. The threshold of three is the smallest strict majority, since a lower one would admit structure two models could jointly hallucinate, whereas a higher one would discard too much (41.5% of argument-bearing utterances are flagged by all five, 57.6% by at least four, and 71.7% by at least three). The result is a consensus graph of 726 argument nodes (491 claims, 235 premises) and 318 relations (229 support, 89 attack) across the six events; individual agents agree with the consensus at detection F_1 between 0.86 and 0.91. It is worth stressing that the consensus graph, and not any single agent's output, is the input to entity linking and digest generation, so that the reliability signal does real work in the pipeline rather than merely decorating it.

5.3. Entity linking

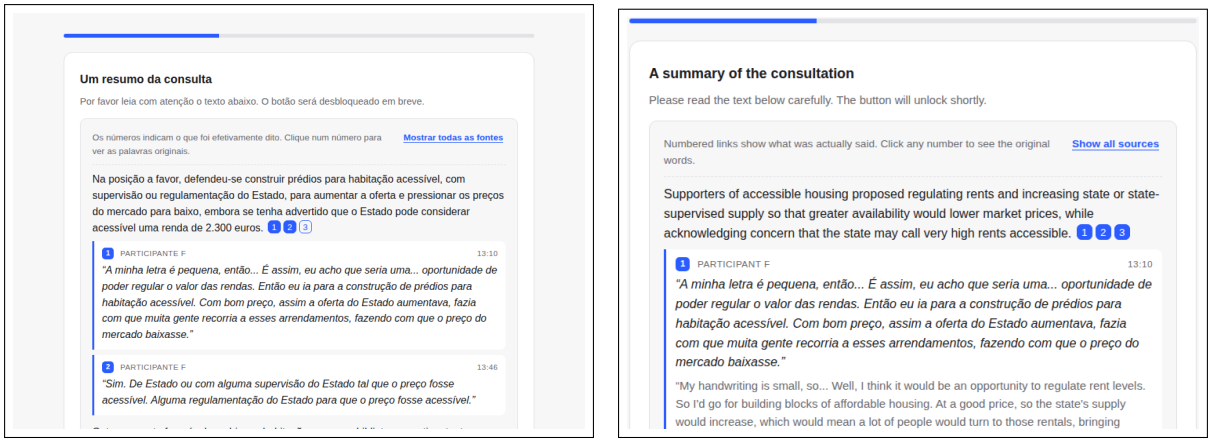
Entity mentions are detected over the source utterances of consensus nodes by an LLM prompted in Portuguese to mark persons, organisations, locations, events and Lisbon-specific *civic* entities. We use an LLM rather than an off-the-shelf tagger precisely because that last category is exactly what standard NER misses. Each surface form (the verbatim, literal textual mention as transcribed from spoken dialogue prior to normalisation or lemmatisation) is then linked against two knowledge bases that fail differently: DBpedia Spotlight [35] disambiguates prominent entities in context, whereas a Wikidata [37] entity search catches the local long tail similarity, in a deliberately precision-first setting, as a wrong anchor poisons every digest built on it, while a missed link merely lands on the worklist. Table 2 summarises the outcome: 81.3% of surface forms link to at least one base, and over half of the 188 entities are anchored in both.

The 60 recurring surfaces that resolve to neither base form a candidate worklist (a structured queue of unlinked candidate entities compiled for systematic manual inspection, disambiguation, and knowledge-base curation). or the *cartão navegante* transit pass) sit alongside ASR-corrupted variants of entities that are in fact present (*Alcoxete* for Alcochete, or *aeroporto Luís de Camarões* for the Humberto Delgado airport). We therefore report the worklist as a curation input rather than a completed KB-completion contribution. Moreover, it exhibits a compounding-error pattern (i.e. ASR error propagating into linking failure) that any pipeline over spoken civic data must budget for.

Table 2
Left: entity-linking coverage over the consensus graph. Right: consensus-graph statistics per event (the 188 entities are corpus-unique, so no per-event split is given).

Quantity	Value
Mentions (LLM)	594
Unique surface forms	332
Surfaces linked	270 (81.3%)
Unique entities	188
via Spotlight	111 (59.0%)
via Wikidata QID	184 (97.9%)
in both	107 (56.9%)
Unlinked worklist	60

Event	Utt.	Claims	Prem.	Rel.
DFG1	456	96	53	76
DFG2	324	76	18	43
DFG3	490	90	50	70
DFG4	323	72	55	50
DFG5	312	93	37	50
DFG6	272	64	22	29
Total	2,177	491	235	318



(a) Portuguese, as served to PT readers

(b) English, as served to EN readers

Figure 2: The graph-grounded digest (C1) as served in the pilot's GRAPH condition, shown side by side in (a) the Portuguese and (b) the English study. Each digest sentence ends in numbered citations that open the verbatim *source* utterance with speaker pseudonym and timestamp; here the first sentence's citations are expanded. Note that, in (b), the digest prose is English while its citations still resolve to the original Portuguese speech, illustrating the pipeline's cross-lingual traceability.

6. Narrative Digests

From each event's consensus graph we generate a narrative digest (mean 573 words, in both Portuguese and English) that retells the deliberation as prose under two constraints: *balance* (proportional coverage of stance clusters) and *traceability* (sentence-level provenance pointers).

Cross-lingual generation and translation. To preserve the colloquial authenticity of the dialogue and prevent compounding errors, all upstream extraction (transcription, multi-agent mining, and entity linking) runs strictly in native European Portuguese. For English digests, rather than translating noisy transcripts, generation is conditioned directly on the structured Portuguese consensus graph. The generator receives the Portuguese proposition nodes together with English label descriptions retrieved via linked DBpedia/Wikidata URIs; it produces fluent English narrative prose while retaining the original citation IDs pointing back to the verbatim Portuguese source utterances (illustrated in Figure 2b). Table 3 illustrates this alignment: an AIF consensus node mined from Portuguese speech is verbalised into Portuguese and English digest sentences, both preserving the exact provenance key resolving to the original speaker's utterance.

Figure 2 shows a graph-grounded digest as readers encountered it in the polling pilot, in both language versions: each sentence ends in numbered citations that expand to the verbatim Portuguese utterance, together with the speaker pseudonym and timestamp.

Table 3
Illustrative cross-lingual alignment: from an AIF consensus graph node (mined in Portuguese) to generated Portuguese (PT) and English (EN) digest sentences, both linked to the verbatim source utterance.

Component	Content / Representation
Source utterance	`DFG1_u0084` (Speaker P3, 14:22): *"Se o aeroporto for para Alcochete, as pontes e os comboios não aguentam o tráfego diário..."*
Consensus node	`I-node:e1_c014` | Type: Claim | Stance: Con | Target: *Novo Aeroporto*
Proposition (PT)	*"Alcochete tem acessos rodoviários e ferroviários insuficientes para suportar o novo aeroporto."*
Entity URI	`dbr:Alcochete` (`wikidata:Q818581`)
Generated PT	Alguns participantes alertaram que Alcochete tem acessos insuficientes de transporte para o aeroporto [1].
Generated EN	Several participants cautioned that Alcochete lacks adequate transport infrastructure for the airport [1].
Traceability	Citation [1] resolves cross-lingually to verbatim Portuguese utterance `DFG1_u0084`.

To test the benefit of grounding on the generation side, we compare two conditions over the same events: (i) C1, *graph-grounded*, generated from the consensus graph under the balance and provenance constraints; and (ii) C2, *transcript-direct*, a straightforward LLM summary of the raw transcript with no graph and no citations (prompts in Appendix A). Each digest sentence is judged for directional entailment against its evidence, following the natural language inference (NLI) paradigm of faithfulness [27, 28]. Specifically, each digest sentence is evaluated as an NLI *hypothesis* (H). For C1 (graph-grounded), the *premise* (P) consists of the verbatim Portuguese source utterances retrieved via the sentence's explicit citation keys. For C2 (transcript-direct), which carries no citations, P consists of the top-k ($k = 3$) transcript utterances retrieved by lexical BM25 overlap against H. An independent LLM judge (`gemini-3.1-pro-preview`, deliberately chosen from a different family than the generator `gpt-5.6-luna` to prevent self-preference bias [33]) evaluates whether P logically entails H under a 3-way rubric: (i) *Entailed*: every assertion in H is directly substantiated by P; (ii) *Contradiction*: H directly conflicts with P; (iii) *Unsupported*: H introduces extraneous facts, speculations, or attributions absent from P. We enforce a conservative decision rule: any sentence flagged with ungrounded content, ambiguity, or API evaluation failure strictly defaults to *unsupported*. We note the absence of a human corroboration pass as a limitation in Section 9. Across the six events, 18.5% of graph-grounded sentences were judged unsupported against their cited evidence, versus 79.5% of transcript-direct sentences against retrieved evidence. However, this gap should not be read as a direct faithfulness comparison: C1 sentences are checked against their explicit citations, whereas C2 sentences rely on lexical retrieval, which often fails to find the relevant source text. The 79.5% figure primarily reflects *unverifiability* rather than proven hallucination. The key takeaway is that graph grounding makes verification possible in the first place, whereas unsourced summaries cannot be checked against what participants actually said.

7. Argument-Grounded Polling: Protocol and Pilot

The polling design is between-subjects with three conditions and one stimulus per respondent: (i) Control (neutral one-to-two-sentence descriptions of the three focal proposals, with no argumentation), (ii) Direct (a transcript-direct digest) and (iii) Graph (a graph-grounded digest served with its visible citation interface, as in Figure 2), the digest conditions being C2 and C1 above, generated from the same event.

We adopt a between-subjects design (where each respondent is randomly allocated to exactly one condition and evaluates a single stimulus) to prevent cross-condition contamination: exposing the same participant to both bare proposals and rich narrative digests would compromise the validity of subsequent argument recognition and attitude change measures. Graph differs from Direct in both generation and presentation, which is a deliberate choice - it is, after all, the artifact the protocol proposes serving - whose cost to interpretability we return to below. The design called for the two digests to be matched for length ($\pm 10\%$) and reading level; this check, however, was not enforced at generation time, and the English pair violated it substantially, which in turn affects the English results

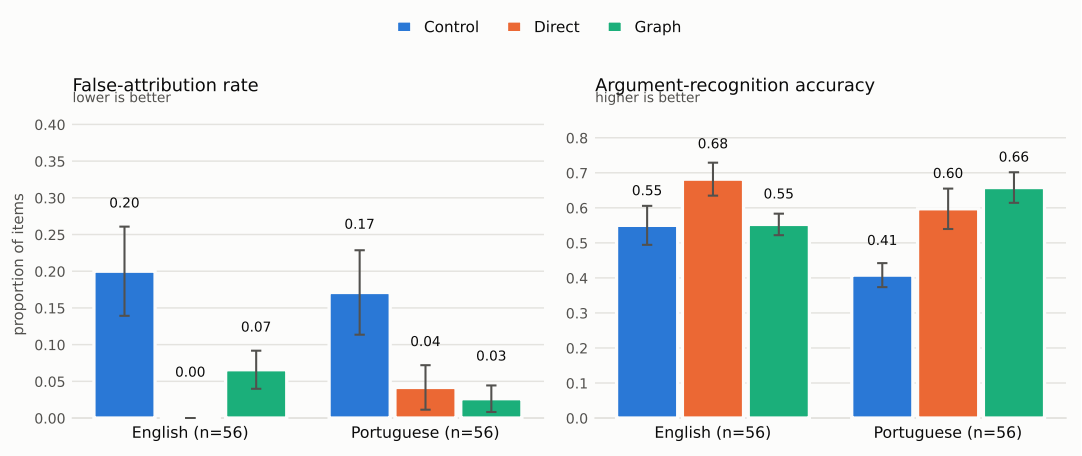


Figure 3: Pilot results ($N = 112$). Serving argumentation - in either digest form - lowers false attribution and raises recognition accuracy relative to the no-digest control. The digest conditions do not separate, except in English, where the graph-grounded stimulus was shorter and less balanced than intended. Bars are means, and whiskers ± 1 SE (Standard Error of the mean).

(Section 9).

Respondents - a fresh crowdworker panel with no connection to the Lisbon events - complete the polling protocol online.

Crowdworker recruitment and experimental procedure. Participants were recruited via Prolific Academic across two language-targeted cohorts: English respondents (filtered for UK/US nationality and English primary language) and Portuguese respondents (filtered for Portuguese nationality and European Portuguese native language). To ensure data quality, eligibility required a minimum platform approval rating of $\geq 98\%$ across at least 50 prior submissions. The study required a median completion time of 10.5 minutes and paid £1.75, yielding an effective hourly rate of ≈£10.00/hr, adhering to UK ethical living-wage standards. Respondents were briefed that they were participating in an academic study evaluating public opinions on municipal urban planning. Before seeing the stimulus, respondents completed a baseline pre-poll of support items across the focal proposals on a 1–7 Likert scale. Next, respondents received their assigned stimulus under an enforced minimum dwell time (60 seconds for digests, 30 seconds for control) to prevent skimming, followed by a brief distractor task. Post-exposure, respondents completed the post-poll support items, a self-reported informedness rating, and the argument-recognition battery. Attention checks (e.g., instructing respondents to select a specific scale option) and minimum time filters were applied; out of 120 recruited submissions, 112 survived the pre-registered screening, yielding 15/22/19 (EN) and 19/18/19 (PT) across CONTROL/DIRECT/GRAPH.

The central measure is an argument-recognition battery of eight statements ("did participants argue this? yes/no/unsure"): four real arguments drawn from the consensus graph and four *foils* (distractor statements representing unvoiced positions, designed to benchmark false attribution and acquiescent guessing; specifically, two *plausible-but-absent* policy ideas verified absent from all six transcripts, and two *stance-reversed* versions of real arguments). In this design, attitude change tests H1, recognition accuracy on the real items tests H2, and the false-attribution rate - the "yes" responses to foils - tests H3, which is the hypothesis the paper most cares about. Perceived balance and trust, demographics and a debrief close the session; the disclosure that the digest was AI-generated comes after the measures, since prior AI-labelling is a known response deflator and not the object of study. Analysis is non-parametric (Mann–Whitney with Cliff's δ), pre-committed, and analysis-blind to condition labels; at ≈19 per cell per language, the pilot estimates rather than confirms. No sensitive personal data were collected; crowdworker identifiers are stored separately from responses under a GDPR consent notice, and all stimuli derive from the anonymised transcripts.

Figure 3 and Table 4 carry the findings. We conclude that H1 is supported: digest readers moved

Table 4
Pilot: Mann–Whitney tests with Cliff's δ. Positive δ favours the first-named group on that measure.

Contrast	Measure	p	δ
Serving argumentation (digests vs. control)			
Digest > Control	attitude change	.010	$+0.30$
Digest < Control	false attribution	<.001	$+0.37$
Digest > Control	recognition acc.	<.001	$+0.41$
Grounding mechanism (Graph vs. Direct)			
pooled	recognition acc.	.23	-0.15
pooled	false attribution	.077	$+0.13$
EN	recognition acc.	.028	-0.37
EN	false attribution	.012	$+0.26$
PT	recognition acc.	.59	$+0.10$
PT	false attribution	.93	-0.01

their attitudes more than controls ($\delta = +0.30$, $p = .010$), driven by the Portuguese sample, albeit with one caveat - digest readers also spent several times longer on their stimulus, so content and exposure duration are confounded by design, and H1 therefore reads as "the package of serving argumentation moves readers" rather than as isolating content from attention. The strongest effects, however, are ones we had treated as a baseline contrast rather than as a hypothesis: readers given *either* digest attributed positions nobody held far less often than controls ($\delta = +0.37$, $p < .001$) and recognised real arguments more accurately ($\delta = +0.41$, $p < .001$). In other words, exposure to the mined argumentation, and not the manner of its writing, is what made readers accurate about the deliberation. By contrast, H2 and H3 are not supported: pooled, graph grounding improved neither recognition nor false attribution; in Portuguese both contrasts are flat, whereas in English both are significant *against* the hypothesis (Table 4), with the confounded English stimuli being the proximate suspect. Perceived balance and trust, finally, did not differ between digest conditions.

Provenance was used, but it did not pay off as an accuracy aid: 55% of Graph readers opened at least one citation, with a strongly bimodal use - many opened none, and a few opened dozens - yet the number opened predicted neither measure. Visible traceability thus attracted half the readers and, at this scale, had no measurable effect. This is a finding about the limits of provenance-as-interface, and one that does not undercut provenance as an *accountability* property, since an auditable record retains its value even when few readers actually audit it.

The protocol, finally, is only as legitimate as the material it serves, so we state the guarantees a pipeline must meet, each supplied by an earlier component (Figure 1): (i) *provenance*, i.e. every served sentence tracing to source utterances; (ii) *balance*, i.e. stance clusters represented in graph proportion so that minority positions survive the telling; (iii) *argument fidelity*, i.e. served structure consensus-backed across independent extractors rather than a single model's artifact; and (iv) *entity fidelity*, i.e. named entities resolving to knowledge-base identities rather than free-text guesses. A poll that meets these transmits the deliberation's actual arguments in proportion and without fabrication, whereas one that does not may leave readers informed, but not deliberative. Our pilot, moreover, indicates which guarantee did observable work for readers. The Control condition was *informed* in the ordinary sense - it described the same three proposals accurately - and yet its readers endorsed invented positions at several times the rate of digest readers. We therefore conclude that the deliberative/informed distinction is empirically visible at pilot scale, and that it turned on *argument fidelity*. What our data do not support, however, is the stronger claim that the grounding mechanism must be a structured graph for this to hold, as a competent transcript summary transmitted the arguments just as well as the graph-grounded digest did. Grounding, then, earns its place through provenance and auditability - properties the reader-side measures were not designed to detect, and which the unsourced condition simply cannot offer - rather than through a measurable accuracy premium for readers.

8. Discussion

The flat results on H2 and H3 require a realistic interpretation. Small sample sizes mean our pilot could only detect large effects, so the lack of difference in Portuguese - where texts were properly matched - suggests any benefit is modest rather than non-existent. This is likely attributed by: (i) short 500-word digests force both summarisation methods to pick the same salient points; (ii) graph grounding's true strengths may only emerge in complex, multi-event debates or controversial topics where minority views risk being lost; or (iii) the effect size is small and requires $N \approx 400$ to observe. Future studies can resolve this by strictly matching summary length, testing graph-grounded text without citation UIs to isolate text from design, and evaluating larger deliberative datasets.

Our results reposition the graph's value rather than diminish it. The graph made our evaluation possible in the first place: real recognition items and absent foils were constructed from it, balance was verified against it, and every sentence remains traceable through it. In a civic deployment, these are vital accountability features owed to third parties-specifically the participants whose words are retold and anyone auditing the results - rather than mere conveniences for readers. The fact that half of readers clicked citations shows that the interface invites verification; that citation clicks did not boost accuracy demonstrates that auditing a source and comprehending a summary are distinct activities. After all, published election records do not automatically make voters better informed, yet no one considers audit trails dispensable.

These properties matter because the framework is designed to scale out. Adding respondents incurs essentially zero marginal cost, while processing a new event requires only transcription, five mining runs, and one generation pass - taking hours instead of the months needed for a standard Deliberative Poll. Furthermore, a single consensus graph can generate digests at various lengths or reading levels without violating stance balance, since proportionality is enforced structurally. It can also bridge languages, as demonstrated by serving Portuguese discussions to English readers via entity-linked referents-though cross-lingual faithfulness warrants dedicated evaluation [32, 39]. Finally, graphs sharing entity anchors can be merged into a cumulative argument record for an entire debate-a scenario where structural balance guarantees become indispensable and unstructured summarisers fail to maintain fair proportionality.

If larger studies confirm our exposure findings, this method offers a practical middle ground between traditional surveys and citizen assemblies: standard polls capture top-of-mind reactions, Deliberative Polls measure informed opinion but cannot scale, and argument-grounded polling captures informed opinion at low cost with sentence-level provenance. For democratic institutions, this helps resolve the transmission bottleneck [3, 4] by turning focus group discussions into reusable digital infrastructure for consultation platforms, participatory budgeting, or civic journalism. However, the approach carries inherent risks. A poll inherits the biases of its source room: (i) an unrepresentative focus group creates an unrepresentative argument pool; (ii) automated extraction might inadvertently mute disfluent or dialectal speech, as seen in our ASR-to-linking error cascade; and (iii) persuasive presentation could be exploited to serve unfaithful summaries. Establishing independent auditing checks for these risks is essential for real-world deployment. Consequently, key next steps include: conducting human audits of consensus-graph nodes to validate semantic accuracy; integrating in-room reaction signals (from Democratic Reflection) to prioritize impactful arguments; curating the unlinked entity worklist for knowledge base completion; and comparing narrative text against interactive visual dashboards [14, 15].

9. Limitations

Without an exhaustive human-coded gold standard across the full corpus, our inter-agent agreement primarily demonstrates *reliability* across diverse LLM families. While our targeted expert spot-check confirmed high human-verified precision for consensus argument nodes (88.0%), relation extraction exhibited lower human agreement (56.0%), mirroring the low model-side $\alpha = 0.08$. Correlated inductive biases across LLM families cannot be entirely ruled out, and relation extraction on colloquial speech

remains a key technical frontier. The reversal in the English trial stems from an unexpected stimulus confound. Because digest generation used a soft length target rather than a strict constraint, the English graph-grounded summary came out at only 279 words (versus 528 words for the direct summary) and omitted opposing and neutral viewpoints. English Graph readers therefore saw roughly half as much text covering fewer perspectives. In contrast, the Portuguese study used length-matched texts (588 vs. 500 words) covering all stance clusters, yielding flat results. This indicates that the English reversal was caused by content imbalance rather than the grounding mechanism itself. Ironically, this failure reinforces our main thesis: unconstrained generation silently violated structural balance, and only the argument graph made this flaw detectable.

Several measurement limits also apply: (i) Graph combined graph-generated prose with a visible citation UI, leaving text generation and UI presentation confounded; (ii) stance-reversed foils exhibited a strong floor effect (only 5 of 112 respondents endorsed them), meaning false attribution was driven almost entirely by plausible-but-absent arguments; (iii) generation-side faithfulness relied on an automated LLM judge without human verification; (iv) Portuguese respondents reported lower prior debate familiarity than English respondents (median 3 vs. 4), meaning expected local context gains did not materialize; and (v) extraction quality on colloquial Portuguese remains the primary technical bottleneck. Window-level extraction is coarse, ASR and diarisation errors propagate into entity linking, relation agreement was near chance ($\alpha = 0.08$), and small cell sizes ($\approx$19 per condition per language) make our pilot exploratory rather than confirmatory.

10. Conclusion

We have shown how the argumentative content of small, expensive deliberative events can be mined into an entity-linked argument graph whose reliability is quantified rather than assumed, re-served as traceable narrative digests, and used as the basis of an argument-grounded polling protocol that extends the deliberative signal beyond the room. Overall, serving that material to fresh readers helped: against a no-digest control, digest readers recognised what was argued more accurately and endorsed invented positions less often, in both languages. However, the mechanism we expected to carry that effect did not separate, as graph-grounded and transcript-direct digests were indistinguishable at reader level, and the single significant reversal traces back to a stimulus we failed to length- and coverage-match.

We take the lesson to be a relocation rather than a retraction. The deliberative/informed distinction is empirically visible - in the gap between readers given the arguments and readers given only the proposals. However, at this scale, reader accuracy turns on serving argumentation at all, whereas graph grounding uniquely provides system-level guarantees: sentence-level auditability, verifiable balance enforcement, and entity anchoring. Indeed, our own English balance failure was only detectable because the graph existed to check against. For a poll to claim deliberative legitimacy, therefore, the graph record remains essential - not as a reader - side accuracy booster, but as the auditable measure that prevents unverified hallucination.

Acknowledgments

This work was carried out within the MultiPoD [3] project, funded by the European Union's Horizon Europe programme under Grant Agreement No. 101178821. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission.

[3] https://multipod-project.eu/

Declaration on Generative AI

The authors used large language models as an integral, disclosed part of the research method: LLM agents perform the argument mining, entity-mention extraction, digest generation and automatic faithfulness judging of Sections 5–6. Argument mining ran five models via OpenRouter (`openai/gpt-5.6-luna`, `qwen/qwen3.7-flash`, `google/gemini-3.1-pro-preview`, `anthropic/claude-opus-4.8`, `z-ai/glm-5.2`); entity-mention extraction used `anthropic/claude-opus-4.8`, digest generation `openai/gpt-5.6-luna`, and faithfulness judging `google/gemini-3.1-pro-preview` (a different family from the generator). All analysis, claims, and this manuscript's final content are the authors' responsibility.

References

[1] J. S. Fishkin, R. C. Luskin, Experimenting with a democratic ideal: Deliberative polling and public opinion, Acta Politica 40 (2005) 284–298. doi:10.1057/palgrave.ap.5500121.

[2] J. S. Fishkin, Democracy When the People Are Thinking: Revitalizing Our Politics through Public Deliberation, Oxford University Press, 2018.

[3] R. E. Goodin, J. S. Dryzek, Deliberative impacts: The macro-political uptake of mini-publics, Politics & Society 34 (2006) 219–244. doi:10.1177/0032329206288152.

[4] OECD, Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave, OECD Publishing, Paris, 2020. doi:10.1787/339306da-en.

[5] M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, M. Botvinick, C. Summerfield, AI can help humans find common ground in democratic deliberation, Science 386 (2024) eadq2852. doi:10.1126/science.adq2852.

[6] A. De Liddo, B. Plüss, A. Ardito, Democratic reflection: Nudging citizens' democratic engagement with political election debates, in: Conference Companion Publication of the 2020 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '20 Companion), 2020, pp. 5–8. doi:10.1145/3406865.3418562.

[7] L. Anastasiou, A. De Liddo, BCause: Reducing group bias and promoting cohesive discussion in online deliberation processes through a simple and engaging online deliberation tool, in: Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023), 2023, pp. 39–49. doi:10.18653/v1/2023.sicon-1.5.

[8] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2019) 765–818. doi:10.1162/coli_a_00364.

[9] S. Elguendouze, L. Anastasiou, E. Hain, E. Cabrio, A. De Liddo, S. Villata, AM4DSP: Argumentation mining in structured decentralized discussion platforms for deliberative democracy, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2025), 2025, pp. 796–805. doi:10.18653/v1/2025.emnlp-demos.61.

[10] C. Chesñevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, S. Willmott, Towards an argument interchange format, The Knowledge Engineering Review 21 (2006) 293–316. doi:10.1017/S0269888906001044.

[11] F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023) e2305016120. doi:10.1073/pnas.2305016120.

[12] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, 2nd ed., Sage, 2004.

[13] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Computational Linguistics 34 (2008) 555–596. doi:10.1162/coli.07-034-R2.

[14] E. Segel, J. Heer, Narrative visualization: Telling stories with data, IEEE Transactions on Visualization and Computer Graphics 16 (2010) 1139–1148. doi:10.1109/TVCG.2010.179.

[15] N. H. Riche, C. Hurter, N. Diakopoulos, S. Carpendale (Eds.), Data-Driven Storytelling, A K Peters/CRC Press, 2018. doi:10.1201/9781315281575.

[16] M. S. Islam, M. T. R. Laskar, M. R. Parvez, E. Hoque, S. Joty, DataNarrative: Automated data-driven storytelling with visualizations and texts, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 19253–19286. doi:10.18653/v1/2024.emnlp-main.1073.

[17] H. Li, Y. Wang, H. Qu, Reflection on data storytelling tools in the generative AI era from the human-AI collaboration perspective, arXiv preprint arXiv:2503.02631 (2025).

[18] H. W. Wang, L. Birnbaum, V. Setlur, Jupybara: Operationalizing a design space for actionable data analysis and storytelling with LLMs, arXiv preprint arXiv:2501.16661 (2025).

[19] Z. Pan, A. Andronis, E. Hayek, O. A. P. Wilkinson, I. Lasy, A. Parry, G. Gadney, T. J. Smith, M. Grierson, Guiding generative storytelling with knowledge graphs, International Journal of Human–Computer Interaction (2025). doi:10.1080/10447318.2025.2603634.

[20] Y. Yang, T. Deng, M.-Y. Kan, KAHAN: Knowledge-augmented hierarchical analysis and narration for financial data narration, arXiv preprint arXiv:2509.17037 (2025).

[21] F. Pianzola, L. Cheng, F. Pannach, X. Yang, L. Scotti, The GOLEM ontology for narrative and fiction, Humanities 14 (2025) 193.

[22] M. Teleki, V. Bengali, X. Dong, S. T. Janjur, H. Liu, T. Liu, C. Wang, T. Liu, Y. Zhang, F. Shipman, J. Caverlee, A survey on LLMs for story generation, Findings of the Association for Computational Linguistics: EMNLP 2025 (2025).

[23] D. Y. Liu, A. Joshi, P. Dawson, Narrative theory-driven LLM methods for automatic story generation and understanding: A survey, arXiv preprint arXiv:2602.15851 (2026).

[24] F. Huot, R. K. Amplayo, J. Palomaki, A. S. Jakobovits, E. Clark, M. Lapata, Agents' room: Narrative generation through multi-step collaboration, in: Proceedings of the 13th International Conference on Learning Representations (ICLR 2025), 2025. ArXiv:2410.02603.

[25] H. Bian, LLM-empowered knowledge graph construction: A survey, arXiv preprint arXiv:2510.20345 (2025).

[26] A. Schimmenti, V. Pasqual, F. Vitali, M. van Erp, Knowledge graphs generation from cultural heritage texts: Combining LLMs and ontological engineering for scholarly debates, arXiv preprint arXiv:2511.10354 (2025).

[27] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

[28] S. Lamsiyah, A. Nourbakhsh, C. Schommer, Trust but verify: A comprehensive survey of faithfulness evaluation methods in abstractive text summarization, in: Proceedings of Recent Advances in Natural Language Processing (RANLP 2025), 2025.

[29] S. Duong, F. Le Bronnec, A. Allauzen, V. Guigue, A. Lumbreras, L. Soulier, P. Gallinari, SCOPE: A self-supervised framework for improving faithfulness in conditional text generation, in: Proceedings of the 13th International Conference on Learning Representations (ICLR 2025), 2025. ArXiv:2502.13674.

[30] H. Sansford, N. Richardson, H. Petric Maretic, J. Nait Saada, GraphEval: A knowledge-graph based LLM hallucination evaluation framework, arXiv preprint arXiv:2407.10793 (2024).

[31] E. Lavrinovics, R. Biswas, K. Hose, J. Bjerva, MultiHal: Multilingual dataset for knowledge-graph grounded evaluation of LLM hallucinations, arXiv preprint arXiv:2505.14101 (2025).

[32] A. Haussmann, The challenge of achieving attributability in multilingual table-to-text generation with question-answer blueprints, arXiv preprint arXiv:2503.23204 (2025).

[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track, 2023.

[34] I. Kartáč, M. Lango, O. Dušek, OpeNLGauge: An explainable metric for NLG evaluation with open-weights LLMs, in: Proceedings of the 18th International Natural Language Generation Conference (INLG 2025), 2025. ArXiv:2503.11858.

[35] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia spotlight: Shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics), 2011, pp. 1–8. doi:10.1145/2063518.2063519.

[36] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia — a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web 6 (2015) 167–195. doi:10.3233/SW-140134.

[37] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85. doi:10.1145/2629489.

[38] M. Boscariol, L. Bulla, L. Draetta, B. Fiumanò, E. Lenzi, L. Piano, Evaluation of LLMs on long-tail entity linking in historical documents, arXiv preprint arXiv:2505.03473 (2025).

[39] L. Draetta, M. Oliverio, V. Ramón-Ferrer, P. F. Balestrucci, F. Corallo, C. Badenes-Olmedo, A. Mazzei, M. A. Stranisci, R. Damiano, TailNLG: A multilingual benchmark addressing verbalization of long-tail entities, arXiv preprint arXiv:2603.27768 (2026).

[40] L. Anastasiou, A. De Liddo, Making sense of online discussions: Can automated reports help?, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21), 2021, pp. 1–7.

doi:10.1145/3411763.3451815.

A. Prompts

All pipeline prompts are Portuguese-native, matching the corpus language; we reproduce them here abridged (JSON output schemas shortened to their first element; full versions in the repository). All extraction and judging stages operate at temperature $T = 0.0$ ($top_p = 1.0$) to minimize generation variance, while digest generation uses $T = 0.3$. Argument mining uses a window size of $W = 15$ utterances with a stride of $S = 10$ utterances (33% overlap).

A.1. Argument mining (per window; all five agents)

```
SYSTEM: És um analista de argumentação especializado em deliberação cívica em
português europeu. A tua tarefa é extrair a estrutura argumentativa de excertos
de grupos de discussão sobre política urbana em Lisboa (transcrições
automáticas de fala espontânea, com ruído). Trabalha SEMPRE em português.
Baseia-te apenas no que é dito no excerto; não inventes conteúdo nem posições
que ninguém defendeu. Se um enunciado não contém argumentação (saudações,
facilitação, hesitações, perguntas do moderador), classifica-o como "other".
Devolve EXCLUSIVAMENTE um objeto JSON válido.

USER: Contexto do tópico (enquadramento do moderador): <topic hint>
EXCERTO DA DELIBERAÇÃO:
[<utterance_id>] <speaker>: <text>   ... (window of utterances) ...

Analisa o excerto acima. Extrai:
1) "units": as unidades argumentativas. Para cada uma: "id" (local, ex. "a1");
   "utterance_ids" (usa exatamente os ids mostrados; nunca inventes ids);
   "unit_type": "claim" | "premise" | "other"; "proposition": reformulação
   curta e fiel (1 frase), sem acrescentar informação; "stance": "pro" | "con"
   | "neutral" | "na"; "target_proposal": proposta visada (ex. "novo
   aeroporto"), "" se não se aplicar.
2) "relations": ligações DENTRO deste excerto: {"from", "to",
   "type": "support" | "attack"}.
Regras: não classifiques enunciados de facilitação como argumentos; mantém
"proposition" ancorada no que foi dito; preferir menos unidades e fiáveis a
muitas e especulativas.
Formato: {"units": [{"id": "a1", "utterance_ids": ["e6_u0003"], "unit_type":
"claim", "proposition": "...", "stance": "con", "target_proposal": "novo
aeroporto"}, ...], "relations": [{"from": "a2", "to": "a1",
"type": "support"}, ...]}
```

A.2. Entity-mention extraction

```
SYSTEM: És um extractor de entidades para deliberação cívica em português
europeu. Recebes enunciados de grupos de discussão sobre política urbana em
Lisboa. Identifica menções a entidades nomeadas e cívicas. Não inventes
entidades que não estejam no texto. Devolve EXCLUSIVAMENTE JSON válido.

USER: ENUNCIADOS:
[<utterance_id>] <text>   ... (batch of 15) ...

Para cada enunciado, extrai as menções a entidades. Tipos: "person";
"organisation"; "location"; "event"; "civic" (entidades cívicas específicas de
Lisboa que provavelmente não estão numa base genérica -- propostas locais,
projetos, instituições municipais, bairros ou equipamentos referidos pelo
nome). Regras: usa a forma de superfície EXATA como aparece no texto; ignora
pronomes e substantivos comuns.
Formato: {"mentions": [{"utterance_id": "e6_u0012", "surface": "Lisboa",
"type": "location"}, ...]}
```

A.3. Digest generation

```
SYSTEM (both conditions): És um redator que reconta deliberações cívicas com
fidelidade absoluta. Não inventas conteúdo. Devolves EXCLUSIVAMENTE JSON
válido.

USER (C1, graph-grounded): Recontas uma deliberação a partir do GRAFO DE
```

```
ARGUMENTOS abaixo. Escreve um resumo narrativo em <língua> de cerca de
<N> palavras que: represente TODAS as posições (a favor, contra, neutro) com
espaço proporcional ao número de afirmações de cada uma; use APENAS o conteúdo
fornecido (não acrescentes factos, nomes ou posições); em cada frase, indica os
IDs dos nós em que se baseia.
GRAFO:
== POSIÇÃO A FAVOR (k afirmações) ==
- AFIRMAÇÃO (<node_id>): <text>
    apoio (<node_id>): <premise text>
    objeção (<node_id>): <counter text>   ... (all stance clusters) ...
Devolve JSON: {"digest": "<texto completo>", "sentences": [{"text": "<frase>",
"cites": ["<id>", ...]}, ...]}.

USER (C2, transcript-direct): Escreve um resumo narrativo em <língua> de cerca
de <N> palavras da seguinte TRANSCRIÇÃO de uma discussão. Sê fiel ao que foi
dito.
TRANSCRIÇÃO:
<speaker>: <text>   ... (full transcript) ...
Devolve JSON: {"digest": "<texto completo>", "sentences": [{"text": "<frase>",
"cites": []}, ...]}.
```