InFACT: Benchmarking LLM Explanations Against Institutional Reasoning for Deliberation-Aware Fact-Checking

Diana Constantina Hoefels
Independent Researcher
Germany
diana.hoefels@gmail.com

Abstract

Explainability in deliberation-support NLP is usually evaluated through post-hoc rationales or model-internal attribution methods, and only rarely against explicit institutional reasoning procedures. We introduce InFACT, a Romanian corpus of professional fact-checking reports that preserves the workflow of editorial epistemic arbitration, namely claim articulation, contextualisation, verification scope, evidence-based verification narrative, and calibrated conclusion. InFACT contains 789 raw reports from *factual.ro* and a processed benchmark release of 788 instances after removal of a singleton non-standard verdict label. Beyond six-way verdict prediction, we position InFACT as a benchmark for LLM explanation alignment, where models must generate short explanations that can be compared directly to gold institutional reasoning. We evaluate InFACT primarily with instruction-tuned LLMs, reporting full-corpus experiments for open-weight models and a matched pilot comparison with GPT-4 Turbo. The resulting evidence shows that verdict prediction and institutional explanation alignment are not the same capability: models that improve verdict accuracy do not necessarily preserve institutional calibration or produce explanations that align with professional verification narratives. These results support the central claim of the paper, namely that InFACT measures not only whether a model reaches a verdict, but also whether it does so in a manner that resembles documented public reasoning.

Keywords: explainable NLP, deliberation technology, institutional reasoning, fact-checking, Romanian, low-resource language, LLM evaluation

1. Introduction

Deliberation is not merely interaction; rather, it is organised epistemic work grounded in the exchange and evaluation of reasons under disagreement (Habermas, 1996; Dryzek, 2000). In public settings, participants and institutions do not simply exchange positions, but formulate claims, delimit what is at issue, bring forward evidence, and calibrate conclusions under conditions of accountability. For NLP, this distinction matters because, if deliberation is modelled only as stance or sentiment, systems may capture disagreement while still failing to represent the procedures through which disagreement is adjudicated (Hautli-Janisz et al., 2024).

Institutional fact-checking makes these procedures unusually explicit. A professional fact-check rarely ends at a binary verdict. Instead, it reconstructs context, specifies verification scope, analyses sources, and issues a conclusion whose force may be partial, conditional, or explicitly bounded. In this sense, institutional fact-checking can be treated as a public record of *deliberative epistemic arbitration*, that is, the process through which an institution translates a contested statement into a documented judgement.

Recent large-scale efforts such as ParlaMint (Erjavec et al., 2024) demonstrate the value of richly structured institutional corpora with extensive metadata and multilingual coverage, although their focus remains parliamentary discourse rather than explicit verification reasoning. Against this background, a resource centred on documented fact-checking procedures addresses a different gap, because it makes institutionally authored justificatory reasoning, rather than deliberative discourse in the broader parliamentary sense, the primary unit of analysis.

This paper introduces InFACT, a Romanian corpus of institutional fact-checking reports derived from *factual.ro* (fac, 2014). The central design choice is to treat the *verification report* as the primary data object rather than merely a claim–label pair. Each record contains the claim and its surrounding context, but also gold institutional reasoning fields, namely `verification_scope`, `verification`, and `conclusion`. In turn, this makes the corpus useful not only for verdict prediction, but also for evaluating whether model-generated explanations resemble the reasoning that institutions make public.

This distinction is especially important for explainability. Prior work has shown that NLP explanations are often assessed through post-hoc rationales, attribution methods, or model-generated justifications whose relation to genuine decision procedures remains uncertain (Lyu et al., 2024; Fragkathoulas and Chlapanis, 2024; Zhao et al., 2023; DeYoung et al., 2020; Jacovi and Goldberg, 2020; Jain and Wallace, 2019; Ribeiro et al., 2016). In InFACT, by contrast, explanation targets are not reconstructed

Proceedings of The 2nd Workshop on Language-driven Deliberation Technology, pages 18–28
May 16, 2026.

after the fact, but embedded in professional verification reports themselves. The resulting benchmark therefore asks not only whether an LLM predicts a verdict correctly, but whether its explanation resembles institutionally grounded public reasoning.

The paper makes four contributions:

1. We release InFACT, a Romanian corpus of structured institutional fact-checking reports, with raw and processed benchmark variants.
2. We document the verdict taxonomy and corpus profile, including label distributions, domains, claimants, temporal coverage, and long-form reasoning structure.
3. We define an explanation-alignment benchmark in which LLMs are evaluated not only on verdict prediction, but also on whether their explanations resemble institutional reasoning and preserve calibrated public judgements.
4. We report full-corpus open-weight LLM results together with a matched pilot comparison against GPT-4 Turbo, showing that verdict competence and institutional explanation alignment can diverge substantially.

The rest of the paper is organised as follows. Section 2 reviews work on deliberation technology, fact-checking, and explanation evaluation. Section 3 introduces the InFACT corpus, including its construction, verdict space, and corpus profile. Section 4 formulates the benchmark task, while Section 5 presents the evaluation framework. Section 6 reports the empirical results, and the remaining sections discuss implications, limitations, ethics, and future work.

2. Related Work

The contribution of InFACT sits at the intersection of three lines of work, namely deliberation technology, fact-checking benchmarks, and explanation evaluation. Reviewing them together clarifies both what the resource inherits from prior work and where it departs from existing benchmark design.

2.1. Deliberation Technology and Public Reasoning

Deliberation theory frames public reasoning as a process of exchanging and assessing reasons in the presence of disagreement (Habermas, 1996; Dryzek, 2000), and in NLP and HCI, this perspective has motivated systems for argument mapping, deliberation-quality assessment, and assistance tools for public discussion (Hautli-Janisz et al., 2024). A recurring limitation, however, concerns evaluation: many deliberation-support systems are assessed on task-specific outputs, whereas standard verification datasets often abstract away the institutional procedures through which evidence is publicly adjudicated. This matters in the present setting because InFACT is intended not only as a fact-checking resource, but also as a benchmark for whether models preserve the justificatory structure of public reasoning.

2.2. Fact-Checking Benchmarks

Fact-checking resources (e.g., LIAR (Wang, 2017), FEVER (Thorne et al., 2018), etc.) have driven progress in supervised claim verification and evidence retrieval. At the same time, the dominant abstraction remains label prediction, so that the workflow that produces an editorial judgement is largely external to the benchmark itself. InFACT complements this landscape by centring the institutional verification narrative as structured gold reasoning text. More specifically, it retains the intermediate fields through which a verdict is justified, thereby making the public reasoning trace itself part of the evaluation object.

2.3. Explainability and Explanation Evaluation

Explainability in NLP has been studied through post-hoc rationales, feature-attribution methods, and model-generated explanations (Lyu et al., 2024; Fragkathoulas and Chlapanis, 2024; Zhao et al., 2023; DeYoung et al., 2020; Jain and Wallace, 2019; Ribeiro et al., 2016). A persistent concern in this literature is *faithfulness*, that is, whether an explanation genuinely supports the prediction rather than merely sounding plausible (Jacovi and Goldberg, 2020). In InFACT, explainability can be evaluated more directly because the corpus contains professional, institutionally accountable explanations.

3. The InFACT Corpus

This section describes how InFACT is constructed and why its structure matters for the benchmark. It first outlines data collection and release format, then describes the verdict space, and finally summarises the corpus properties that motivate the evaluation design.

3.1. Corpus Construction

InFACT is defined as a corpus of *institutional verification records*, as each record preserves the publicly visible stages through which a professional fact-checking organisation evaluates a claim, namely claim articulation, contextualisation, verification scope, evidence-backed verification narrative, and

editorial conclusion. In this setting, the corpus supports deliberation-aware modelling precisely because systems must preserve procedural justification and calibrated outcomes rather than merely output a label.

The dataset is constructed from publicly available fact-checking reports published as structured web articles on *factual.ro* (fac, 2014), a Romanian fact-checking platform operated by Funky Citizens.[1] The platform describes itself as the first independent fact-checking website in Romania. The goal was not to create new annotations, but to preserve the platform's editorial workflow in machine-readable form.

The dataset was compiled by crawling a curated set of report URLs and extracting the following metadata and textual components:

- **Metadata**: verification date, claimant attribution, source outlet, topical domain, and editorial verdict.
- **Procedural segments**: `claim_text`, `context`, `verification_scope`, `verification`, and `conclusion`.

During ingestion, lightweight normalisation is applied, including Unicode normalisation, whitespace clean-up, and removal of duplicated boilerplate spans where present. Romanian diacritics and the substantive content of the reports are preserved, including explicit normative references when cited. Automated validation confirms that the released dataset contains no missing values.

Importantly, the aim is not to treat the source verdicts as philosophically final ground truth, but rather as institutional editorial judgements produced under a documented public verification procedure. According to its published methodology, *factual.ro* focuses on publicly checkable factual claims and grounds its reports in source-based verification using journalistic and research-oriented practices.[2] Accordingly, the object of study in InFACT is not truth in the strongest abstract sense, but institutional reasoning as it is publicly documented, justified, and communicated.

3.2. Corpus Structure and Releases

The corpus is released in two forms. The raw release is a TSV file with 789 reports and 12 fields:

- `record_id`
- `source_url`
- `date_verified`
- `author_claim`
- `source_outlet`
- `claim_text`
- `context`
- `verification_scope`
- `verification`
- `conclusion`
- `domain_claim`
- `verdict_original`

For modelling and analysis, a processed benchmark release is provided that adds `verdict_normalized`, `label_id`, `label_binary`, `claim_len`, and `context_len`. The processed release therefore contains 17 fields in total and, after excluding a singleton non-standard verdict label, 788 instances. To make the structure of the resource more concrete, Appendix A provides a shortened example record showing how a public claim is transformed into an institutional verification record through contextualisation, scope definition, evidence-based reasoning, and editorial conclusion.

3.3. Verdict Space and Normalisation

InFACT preserves the platform's editorial verdict taxonomy in `verdict_original`. In the raw data, eight verdict labels appear: *fals*(false), *adevărat* (true), *trunchiat* (truncated), *parțial adevărat* (partially true), *parțial fals* (partially false), *imposibil de verificat* (impossible to verify), *inexplicabil* (inexplicable), and *numai cu sprijin instituțional* (institutional support only). The Romanian labels are retained to preserve the institutional taxonomy while providing English glosses in the prose. The label *inexplicable* occurs only once.

To reduce sparsity and support benchmarking, the Romanian editorial verdicts are mapped into a six-way normalised space: True, False, Mixed, Mostly True, Mostly False, Unverifiable. This mapping is informed by prior fact-checking benchmarks but is defined in a task-oriented manner for InFACT, preserving distinctions that matter for institutional reasoning, namely partial support, partial refutation, and explicit non-verifiability.

The mapping preserves the epistemic force of the editorial labels while reducing label sparsity. *adevărat* and *fals* map directly to True and False. *parțial adevărat* and *parțial fals* map to Mostly True and Mostly False. *trunchiat* is mapped to Mixed, since it typically marks a claim that is misleading through omission or incomplete framing rather than simply false. *imposibil de verificat* and *numai cu sprijin instituțional* are mapped to Unverifiable, since neither yields a determinate public truth-conditional judgement. The singleton label *inexplicabil* is excluded from the processed benchmark because it does not form a stable learnable class.

Binarisation of labels is also provided by mapping True to 1 and all other normalised categories

[1] https://funky.ong/en/

[2] https://blog.factual.ro/2021/01/20/metodologie/

to 0. This supports strict endorsement experiments but is not the main label space of the paper. Table 1 reports the original and normalised verdict distributions in the processed benchmark release.

Original verdicts (verdict_original)		
Label	**Count**	**%**
fals	287	36.4
adevărat	173	22.0
trunchiat	124	15.7
parţial adevărat	121	15.4
parţial fals	74	9.4
imposibil de verificat	8	1.0
numai cu sprijin instituţional	1	0.1
Total	788	100.0
Normalised verdicts (verdict_normalised)		
Label	**Count**	**%**
False	287	36.4
True	173	22.0
Mixed	124	15.7
Mostly True	121	15.4
Mostly False	74	9.4
Unverifiable	9	1.1
Total	788	100.0

Table 1: Verdict distributions in the processed InFACT benchmark release.

3.4. Corpus Profile

Before introducing the benchmark tasks, this section briefly characterises the processed release. These corpus statistics are not only descriptive, but also analytically informative, as they highlight structural properties that are likely to shape downstream model behaviour, including label imbalance, topical concentration, temporal clustering, and the asymmetry between short public claims and long-form institutional reasoning.

The processed corpus spans 3,197 days, from 2016–05–19 to 2025–02–18, and contains 231 unique claimants. The label distribution is moderately imbalanced: FALSE is the largest class (36.4%) and UNVERIFIABLE the smallest (1.1%), yielding a majority/minority ratio of 31.9×.

In terms of domain representation, InFACT includes an editorial domain label spanning six domains: *Politică* (Politics) (238; 30.2%), *Finanțe* (Finance) (164; 20.8%), *Economie* (Economy) (156; 19.8%), *Justiție* (Justice) (137; 17.4%), *Externe* (Foreign Affairs) (57; 7.2%), and *Energie* (Energy) (36; 4.6%). Domain diversity measured by Shannon entropy is $H = 2.372$ bits, indicating broad but uneven topical coverage. The distribution shows that the resource is concentrated in politically salient domains while still covering multiple policy areas.

Furthermore, the corpus is also temporally uneven. Publication volume peaks in 2019 (238 reports; 30.2%), which reflects political cycles and salient public events. This matters for evaluation because shifts in rhetoric, agenda, and institutional attention may affect generalisation.

A defining property of InFACT is the asymmetry between short claims and long verification narratives. Table 2 summarises word-count profiles for the main procedural fields. In particular, verification has a mean length of 443.3 words, compared with 37.9 words for claim_text. This gap is central to the benchmark: the modelling problem is not only to classify short public statements, but to recover judgements that are produced through much longer institutional reasoning.

Field	MEAN	MEDIAN	MAX
claim_text	37.9	31	214
context	97.3	81	674
verification_scope	20.0	17	114
verification	443.3	344	3,075
conclusion	60.8	50	1,516

Table 2: Word-count profile for procedural segments in the processed InFACT release.

4. Benchmark Formulation

The corpus design makes it possible to define a benchmark that goes beyond verdict classification alone. Rather than treating the task as the prediction of a label in isolation, InFACT requires a model to generate both a verdict and a short explanation, which can then be assessed against institutionally documented reasoning.

For each record i, the model receives:

$$x_i = \langle \texttt{claim\_text}_i, \texttt{context}_i \rangle$$

The model outputs a verdict

$$\hat{y}_i \in \mathcal{Y} = \{\textsc{True}, \textsc{False}, \textsc{Mixed}, \textsc{Mostly True}, \textsc{Mostly False}, \textsc{Unverifiable}\}$$

together with a short explanation $\hat{e}_i$ of 3–5 sentences.

We compare $\hat{e}_i$ to verification as the long-form institutional reasoning reference and to conclusion as the shorter editorial justification reference. In this setting, explanation alignment is framed as alignment with institutional reasoning rather than as a single-reference correctness test. More specifically, the institutional report provides a documented public reasoning trace against which model explanations can be compared, while allowing for the possibility that multiple valid explanations may exist in principle.

Operationally, the evaluation pipeline constructs a separate zero-shot prompt for each benchmark

instance from the claim and its context, queries the model, parses the returned verdict and explanation into a structured representation, and then evaluates the outputs under a shared scoring framework. This design supports reproducible benchmarking across models while enabling large-scale comparison under identical prompting and evaluation conditions.

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM.

5. Evaluation Framework

In line with the paper's LLM-centred contribution, the experimental design is organised around explanation alignment rather than verdict classification alone. At the same time, reference baselines remain useful because they indicate how much of the verdict signal can be recovered without generation and therefore provide context for interpreting the LLM results.

For the LLM experiments, the default input consists of `claim_text` and `context`. This reflects the intended user-facing setting in which a model receives the public claim together with surrounding contextualisation, but not the gold institutional reasoning itself. For some auxiliary verdict baselines, we additionally consider `verification_scope` as an input field in order to measure how much institutional framing contributes to verdict prediction.

The benchmark experiments use instruction-tuned generative models, i.e., open-weight models, such as `llama3:latest-instruct 4.7 GB`, `qwen2.5-7B-instruct`, and `mistral:7b-instruct 4.2 GB`, evaluated on the full processed corpus and a matched `GPT-4 Turbo` pilot. We use two complementary protocols. First, the full-corpus open-weight protocol evaluates open LLMs on all 788 processed instances, which is the primary benchmark setting for InFACT as an explanation-alignment resource. Second, the matched pilot protocol evaluates proprietary and open models on the same 174-instance subset, thereby enabling a direct comparison between GPT-4 Turbo and open-weight models under an identical scoring framework.

For all LLM runs, models are prompted to return a six-way verdict together with a short explanation in Romanian. Evaluation is based on verdict accuracy, macro-F1, nuance collapse, ROUGE-L against `conclusion` and `verification`, and evidence-type overlap.

The main empirical focus of the paper is therefore not general verdict classification, but whether large language models can produce explanations that remain aligned with the institutional reasoning procedures documented in InFACT. Accordingly, non-generative classifiers and encoder models are treated as reference verdict baselines, while the central benchmark analysis concerns explanation alignment in instruction-tuned LLMs.

The benchmark requires two complementary types of evaluation. Verdict prediction is assessed with standard classification metrics, whereas explanation alignment is assessed with overlap and calibration diagnostics designed to capture whether a model preserves institutional nuance. For six-way verdict prediction, we report accuracy, macro-F1, and weighted-F1, with macro-F1 receiving particular attention because the verdict distribution is moderately imbalanced and several categories are substantially less frequent than FALSE.

We also compute a nuance collapse rate (NCR). LLMs often collapse calibrated verdicts into binary TRUE/FALSE predictions. To quantify this behaviour, NCR is defined over the subset of gold labels that encode institutionally nuanced outcomes:

$$\mathcal{Y}_{\text{nuanced}} = \{\textsc{Mixed}, \textsc{Mostly True}, \textsc{Mostly False}, \textsc{Unverifiable}\}.$$

$$\text{NCR} = \frac{\sum_{i=1}^{N} \mathbf{1}[y_i \in \mathcal{Y}_{\text{nuanced}} \wedge \hat{y}_i \in \{\textsc{True}, \textsc{False}\}]}{\sum_{i=1}^{N} \mathbf{1}[y_i \in \mathcal{Y}_{\text{nuanced}}]}.$$

A higher NCR indicates greater loss of deliberative calibration, as the model replaces institutionally nuanced verdicts with simpler binary outcomes.

To operationalise explanation comparison, we use established text-generation metrics, namely ROUGE-L (Lin, 2004) for sequence-level overlap, and report the metric against both `conclusion` and `verification`. The released scripts also support BERTScore (Zhang et al., 2020), but those values are not included in the current evaluation. ROUGE-L is retained as the primary reported overlap metric because it provides a simple and transparent signal of textual alignment with institutional reasoning across all evaluated models.

Further, the purpose of these metrics is not to claim explanatory correctness, but to estimate how closely a model explanation follows institutional reasoning. Importantly, the explanation task is not intended as a single-reference rationale generation benchmark in which any deviation from the institutional text is treated as incorrect reasoning. Rather, it is an alignment benchmark, and overlap-based metrics such as ROUGE-L and evidence-type overlap should therefore be interpreted as indicators of alignment with institutional reasoning rather than as measures of exclusive validity.

As a deliberation-oriented diagnostic, coarse evidence signals are extracted from both the model

explanation and the institutional reference. The aim is not to identify exact factual matches, but to assess whether the model appeals to similar kinds of evidence as those used in institutional reasoning.

Detection is implemented with a rule-based pattern matcher in the released codebase.

The detector relies on keyword lists, abbreviation patterns, numeric expressions, URL-like strings, and date-like expressions rather than on a trained classifier.

A rule-based detector is used over five evidence categories: LAW, STATISTICS, AUTHORITY, SOURCE/URL, and TIME. LAW captures references such as *lege* (law), *art.* (article), *ordonanță* (ordinance), OUG (Government Emergency Ordinance), and *Hotărâre de Guvern (HG)* (Government Decision). STATISTICS captures numerals, percentages, and quantity expressions. AUTHORITY captures institution names and official titles, such as ministries, agencies, Eurostat, INS, and other public authorities. SOURCE/URL captures explicit links or source-like references. TIME captures years and date-like expressions.

Let $E(\cdot)$ map a text to the set of evidence categories detected in that text. Evidence overlap is then defined as the mean Jaccard similarity between the evidence-category set extracted from the model explanation and that extracted from the institutional reference:

$$\text{EvidenceOverlap} = \frac{1}{N}\sum_{i=1}^{N}\frac{|E(\hat{e}_i) \cap E(r_i)|}{|E(\hat{e}_i) \cup E(r_i)|},$$

where r_i denotes either `verification` or `conclusion`. This diagnostic captures whether the model appeals to evidence in a manner that is structurally similar to the institutional reference, even when the wording differs.

6. Results

This section presents the empirical results. The analysis begins with the full-corpus open-weight setting, which most closely reflects the intended use of InFACT as a benchmark, and then turns to a smaller matched comparison that also includes a proprietary model.

Non-generative reference baselines for verdict prediction were also evaluated, including lexical classifiers, input-structure ablations, and encoder-based transformers. These experiments show that verdict prediction is non-trivial and that institutional framing materially improves performance. In particular, the strongest verdict accuracy among the reference baselines is obtained by an SVM using claim, context, and verification scope, while Romanian BERT yields the strongest transformer macro-F1.

Metric	Llama 3	Qwen	Mistral
Acc	**0.262**	0.261	0.253
Macro-F1	0.128	**0.180**	0.117
ROUGE-L$_{con}$	0.146±0.069	0.136±0.065	**0.147±0.079**
ROUGE-L$_{ver}$	0.062±0.039	0.060±0.035	**0.072±0.044**
EvOverlap$_{con}$	0.603±0.406	0.580±0.407	**0.642±0.411**
EvOverlap$_{ver}$	**0.603±0.406**	0.372±0.271	0.415±0.300
Collapse	**0.282**	0.702	0.391

Table 3: Full-corpus open-weight LLM evaluation on Llama 3, Qwen2.5-7B-Instruct, and Mistral 7B-Instruct.

Since the central focus of the present paper is LLM explanation alignment rather than non-generative classification, the detailed baseline tables are reported in the Appendix. These baselines are included as reference points for verdict recoverability rather than as state-of-the-art competitors in generative explanation alignment.

The main benchmark use case for InFACT is full-corpus evaluation of open-weight LLMs. In this setting, models are run on all 788 processed instances and assessed jointly on verdict prediction, nuance preservation, and explanation alignment to `conclusion` and `verification`. This is the primary setting for establishing whether InFACT can differentiate open-weight LLMs not only by verdict competence, but also by the degree to which they reproduce institutional reasoning. Table 3 reports the full-corpus open-weight results.

A first observation is that no single open-weight model dominates all metrics. Llama 3 attains the strongest accuracy (0.262) and the lowest nuance collapse rate (0.282), which suggests a comparatively better balance between coarse verdict assignment and the preservation of calibrated categories. By contrast, Qwen2.5 reaches a very similar accuracy but the strongest macro-F1 (0.180), while also exhibiting by far the highest collapse rate (0.702); taken together, this pattern suggests that better class balance at the verdict level does not necessarily translate into deliberatively calibrated behaviour. Mistral, in turn, is weaker on verdict metrics than Llama 3 and Qwen2.5, yet comparatively stronger on overlap-based alignment measures, especially ROUGE-L against `verification` and evidence overlap against `conclusion`.

What matters here is not simply which model performs best, but the fact that the benchmark separates these dimensions in a meaningful way. In other words, the full-corpus setting already shows that verdict competence, nuance preservation, and institutional explanation alignment should be treated as partially independent properties rather than as interchangeable indicators of model quality.

This distinction is directly relevant to deliberation-aware evaluation, because a model may produce a

Model	Acc	F1	R-L$_{con}$	Collapse
GPT-4 Turbo	0.218	0.181	**0.168±0.061**	0.311
Qwen2.5-7B	**0.264**	**0.196**	0.030±0.064	0.729
Llama 3 8B	0.149	0.124	0.034±0.073	**0.300**
Mistral 7B	0.189	0.065	0.128 ± 0.316	0.442

Table 4: Matched 174-instance LLM comparison. R-L$_{con}$ denotes ROUGE-L against `conclusion`; F1 denotes macro-F1.

plausible verdict while still failing to preserve the calibrated and evidence-structured reasoning through which that verdict is institutionally justified.

To provide a controlled comparison that also includes a proprietary model, we report a matched pilot evaluation on the first 174 processed instances. For compactness, the matched comparison reports conclusion-level ROUGE-L as the primary explanation-overlap metric. The motivation for this choice is that `conclusion` provides a shorter and more directly comparable institutional reference for 3–5 sentence model explanations.

This setting allows direct comparison between GPT-4 Turbo and open-weight LLMs under the same subset and the same evaluation protocol, which in turn makes it possible to distinguish dataset-scale effects from genuine model differences. Table 4 illustrates the matched 174-instance LLM comparison results.

GPT-4 Turbo serves an important role in this comparison. Although its verdict performance is relatively weak, it provides a useful proprietary anchor for the explanation-alignment task because it exhibits a qualitatively different behaviour from the open-weight pilots. More specifically, GPT-4 Turbo attains lower verdict accuracy than Qwen2.5, but aligns much more closely with institutional reasoning and preserves nuanced verdict categories substantially better.

Qwen2.5, by contrast, slightly improves verdict performance on the same subset, but does so at the cost of much poorer calibration and much weaker explanation alignment. Llama 3 shows the weakest verdict performance on the matched subset, although it also yields the lowest collapse rate among the four models. Mistral occupies an intermediate position: its verdict metrics remain weak, yet its conclusion-level overlap is notably stronger than that of the two open models Qwen2.5 and Llama 3. This is exactly the type of divergence the benchmark is designed to detect.

Taken together, the matched subset strengthens the central interpretive claim of the paper. A model may perform relatively better on verdict assignment while aligning substantially worse with institutional reasoning, and the reverse pattern may also hold. In this sense, InFACT does not collapse explainability into verdict accuracy, but instead makes visible the trade-off between prediction and institutionally grounded justification.

7. Discussion

The main contribution of InFACT comes from the fact that it makes the reasoning trace itself available for evaluation. This matters because, in a deliberation-aware setting, the relevant question is not only whether a model reaches the correct category, but also whether it preserves the justificatory structure through which contested claims are publicly adjudicated.

Against this background, the full-corpus open-weight results are already informative. They show that open models can be meaningfully differentiated not only by verdict metrics, but also by their calibration behaviour and by how closely their explanations resemble institutional reasoning. Just as importantly, the results do not point to a single dimension along which all desirable behaviour aligns. Llama 3 is comparatively stronger on accuracy and collapse, Qwen2.5 on macro-F1, and Mistral on several overlap-based explanation metrics. In turn, this suggests that benchmarking institutional reasoning requires a multi-dimensional evaluation framework rather than a single aggregate score.

The matched pilot comparison makes the same point even more clearly. GPT-4 Turbo and Qwen2.5 separate verdict competence from explanation alignment in opposite directions: Qwen2.5 improves slightly on verdict prediction, whereas GPT-4 Turbo aligns more closely with professional justifications and preserves nuance substantially better. Llama 3 and Mistral, meanwhile, occupy yet different positions in this space. Taken together, these patterns are precisely what one would expect from a benchmark that measures alignment with institutional reasoning rather than surface plausibility alone.

At the same time, the corpus itself reveals properties that are relevant for deliberation-aware evaluation. InFACT contains 231 claimants with uneven verdict distributions, and under the strict binary subset several high-profile claimants have false rates above 80%. This creates an obvious shortcut risk: models may learn claimant-specific regularities rather than reasoning over the content of the claim. In a similar vein, a lightweight lexical audit of institutional narratives identifies sparse but non-negligible hedging, certainty, and authority markers, which suggests that calibration is neither reducible to lexical templates nor absent from the data. In this respect, the resource is useful not simply because it contains fact-checks, but because it preserves the kinds of evidential and epistemic cues through which public justification becomes accountable.

Seen in this light, InFACT is relevant to deliberation technology rather than only to fact-checking in the narrow sense. The benchmark does not merely ask whether a model predicts a verdict correctly; it asks whether the model preserves the intermediate reasoning structure through which contested claims are publicly examined, bounded, and justified. In deliberative settings, this distinction matters because users need not only an answer, but also a traceable account of how that answer was reached, what kinds of evidence were considered, and where uncertainty remains.

8. Limitations

Like any institutional resource, InFACT comes with important limitations that shape both its interpretation and its use.

First, the corpus reflects the editorial selection and reasoning conventions of a single institutional platform. It is therefore not a neutral sample of Romanian public discourse, but a sample of what one institution chose to check, how it framed those checks, and how it justified its conclusions.

Second, the explanation-alignment benchmark is stronger than standard post-hoc evaluation, but it is not exhaustive. ROUGE-L and evidence overlap are alignment signals rather than guarantees of evidential validity, and for precisely that reason they should be read as indicators of resemblance to institutional reasoning rather than as measures of exclusive correctness.

Third, although the paper includes full-corpus open-weight evaluation, the matched proprietary comparison is still limited to a 174-instance subset. This provides useful initial evidence, especially for the contrast between verdict competence and explanation alignment, but it does not yet establish a stable benchmark floor or ceiling across the full range of model families that may be relevant. Broader evaluation across proprietary and open models remains a natural next step.

9. Conclusion

We introduced InFACT, a Romanian corpus of institutional fact-checking reports, and positioned it as a benchmark for evaluating LLM explanations against professional institutional reasoning. By preserving verification scope, evidence-backed narratives, and calibrated conclusions, the corpus supports evaluation beyond verdict prediction and makes it possible to assess whether models resemble documented public reasoning.

The current results show that InFACT can separate verdict performance from explanation alignment in a meaningful way. Reference baselines establish that verdict prediction is non-trivial and that structured institutional framing matters. More importantly, the LLM experiments show that stronger verdict prediction does not imply better institutional explanation alignment. In particular, the matched pilot comparison demonstrates that GPT-4 Turbo aligns more closely with professional justifications and preserves nuance better, whereas Qwen2.5 improves slightly on verdict prediction while performing substantially worse on explanation alignment and calibrated verdict preservation. In turn, the full-corpus open-weight experiments show that these trade-offs persist at benchmark scale.

This is precisely where the relevance of InFACT for deliberation technology becomes clearest. In deliberative settings, users often need more than a verdict: they need a trace of how that verdict was reached, what evidence categories were considered, and where uncertainty remains. This matters especially in politically contested discourse, where trust depends not only on correctness, but also on procedural transparency. By preserving verification scope, long-form reasoning, and calibrated conclusions, InFACT supports evaluation scenarios in which a system must justify its output in a form that remains legible to human participants, facilitators, or institutional stakeholders.

Several directions follow naturally from the present results. First, the explanation-alignment benchmark should be extended from the current pilot comparisons to full-corpus evaluation across a broader set of open and closed LLMs. Second, future evaluation should include claimant-disjoint and time-aware splits in order to test robustness under shortcut risk and temporal drift. Third, the current setup can be extended toward evidence-aware generation, where models are asked not only to predict verdicts and explanations, but also to recover verification scope or identify the types of evidence used in institutional reasoning. More broadly, future work should examine whether models can be made not only more accurate, but also more faithful to the calibrated public reasoning that InFACT preserves.

10. Ethical Considerations

Since InFACT concerns political claims and institutional judgements, its use requires particular care. Models trained on InFACT could be misused to make automated credibility judgements about individuals rather than about specific claims in context. For this reason, we recommend reporting uncertainty and avoiding deployments that attribute truthfulness to persons instead of to statements embedded in evidence and time.

At the same time, claimant skew is a real property of institutional fact-checking, but it can also become a shortcut signal for models. This creates a risk that systems learn claimant identity rather than institu-

tional reasoning, which in turn makes attribution-aware analysis and claimant-disjoint evaluation especially important for future work.

11. Data and Code Availability

All texts are derived from publicly available fact-checking reports with source URLs preserved for traceability. We do not add new personal annotations. Instead, we preserve institutional report structure in machine-readable form so that modelling work can be grounded in existing public reasoning practices.

We release the raw TSV, the processed benchmark release, and a reproducibility suite that computes descriptive statistics, verdict baselines, and explanation-alignment diagnostics, including prompt generation and scoring utilities. The repository can be accessed at https://github.com/DianaHoefels/INFACT.

12. Acknowledgements

I would like to thank the editorial team at factual.ro for kindly granting permission to collect and use their fact-checking content for this research. Their cooperation made the construction of the InFACT corpus possible.

13. Bibliographical References

2014. Factual - verificăm faptele. https://www.factual.ro/. Accessed 2024-2026.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

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

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darundefinedis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2024. Parlamint ii: advancing comparable parliamentary corpora across europe: Parlamint ii: advancing comparable parliamentary... *Lang. Resour. Eval.*, 59(3):2071–2102.

Christos Fragkathoulas and Odysseas Spyridon Chlapanis. 2024. Local explanations and self-explanations for assessing faithfulness in black-box llms. In *Proceedings of the 13th Hellenic Conference on Artificial Intelligence*, SETN 2024, page 1–5. ACM.

Jürgen Habermas. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, MA.

Annette Hautli-Janisz, Gabriella Lapesa, Lucas Anastasiou, Valentin Gold, Anna De Liddo, and Chris Reed, editors. 2024. *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey.

14. Language Resource References

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Appendix A. Additional Examples and Baselines

Field	Content (shortened, with English translation)
`claim_text`	"E primul an în care bugetul cercetării crește cu 70%." (*It is the first year in which the research budget increases by 70%.*)
`context`	"În 23 ianuarie 2023, ministrul Cercetării, Inovării și Digitalizării ..." (*On 23 January 2023, the Minister of Research, Innovation and Digitalisation ...*)
`verification_scope`	"Cum a crescut bugetul cercetării în ultimii ani." (*How the research budget has increased in recent years.*)
`verification`	"În bugetul național pentru 2023, pentru cercetare fundamentală și cercetare dezvoltare s-au alocat 3,2 miliarde de lei (credite bugetare) (pagina 77, capitolul 5301) ..." (*In the national budget for 2023, 3.2 billion lei were allocated to fundamental research and research and development (budgetary credits) (page 77, chapter 5301) ...*)
`conclusion`	"Afirmația ministrului este adevărată. Este primul an în care bugetul alocat cercetării crește ..." (*The minister's statement is true. It is the first year in which the budget allocated to research increases ...*)
`verdict_original`	*adevărat* (*true*)
`verdict_normalized`	TRUE

Table 5: Shortened example of an InFACT record illustrating the structure of an institutional fact-checking report on *factual.ro*.

Classifier	Accuracy	Macro-F1	Weighted-F1
Logistic Regression	0.381 ± 0.027	0.267 ± 0.021	0.372 ± 0.027
SVM	0.390 ± 0.035	**0.269 ± 0.030**	0.376 ± 0.036
Naive Bayes	0.396 ± 0.024	0.137 ± 0.024	0.258 ± 0.033
Random Forest	**0.419 ± 0.023**	0.220 ± 0.018	**0.324 ± 0.022**

Table 6: Stratified 5-fold cross-validation for six-way verdict prediction on InFACT using TF–IDF features over `claim_text`.

Input	Accuracy	Macro-F1	Weighted-F1
Claim-only	0.390	0.269	0.376
Claim + Context	0.440	0.310	0.420
Claim + Context + Scope	**0.460**	**0.330**	**0.440**

Input-structure ablation for the SVM lexical baseline.

Model	Accuracy	Macro-F1	Weighted-F1
XLM-RoBERTa (claim+context)	0.371 ± 0.035	0.290 ± 0.015	0.353 ± 0.025
XLM-RoBERTa (claim+context+scope)	0.345 ± 0.018	0.258 ± 0.008	0.325 ± 0.008
Romanian BERT cased v1 (claim+context)	**0.381** ± 0.036	**0.357** ± 0.033	**0.387** ± 0.038
Romanian BERT cased v1 (claim+context+scope)	0.363 ± 0.014	0.332 ± 0.020	0.366 ± 0.013
Romanian BERT uncased v1 (claim+context)	0.300 ± 0.053	0.289 ± 0.052	0.299 ± 0.055
Romanian BERT uncased v1 (claim+context+scope)	0.297 ± 0.055	0.290 ± 0.056	0.291 ± 0.061

Table 7: Transformer baselines for six-way verdict prediction on InFACT under stratified 5-fold cross-validation.