Deterministic-First Fact-Checking of Local News Against Multilingual Sources: A Pre-Registered Evaluation on Held-Out Articles

Alec Meeker, Bushwick Daily, Brooklyn, New York

Manuscript, 28 October 2026

Abstract

Background. Every news article needs a check of each sentence against the documents it was written from before it is published, whether a reporter or a language model wrote it. Model-only checking is costly, varies between runs, and fails silently in ways an editor cannot audit.

System. We built a fact-check desk for a New York City local news publisher that settles what it can with deterministic code and asks a language model only about the rest. It reads an article and the documents behind it, whoever wrote the article. Code holds every evidence document whole, settles sentences with five rules (no checkable fact; a quotation in no document; a figure in no independent document; one passage holds the whole sentence; nothing in any document shares its content), retrieves passages for the remainder, verifies every excerpt the model cites, and re-checks every confirmed sentence for stale relative dates, date ranges, the outlets it credits and durations that contradict the article's own years. The model reads only located passages; its readings are recorded and replayed.

Evaluation. Thirteen pre-registered rounds, each on ten never-checked articles from the publisher's review queue (127 of the 130 drafted by its language-model writer): Part A, forty factual errors planted by a blind writer agent plus ten meaning-preserving rewrites; Part B, every sentence labelled by two blind reader agents against the same evidence, with adjudication and checks outside the evidence. A round passes when at least 36 of 40 planted errors are not confirmed, at least 32 are blocked, rewrites are confirmed, no factual error is confirmed (B1), and at most 2% of the round's sentences are correct but blocked (B2).

Results. Over rounds 2-13 (2,934 sentences), the desk did not confirm 476 of 482 planted errors (98.8%) and confirmed 111 of 113 rewrites. Confirmed factual errors fell from 1-6 per round in rounds 2-7 to 0-2 in rounds 8-13; under the publisher's standard, which does not count hair-splitting or stale verb tense as factual errors, 15 of the 30 confirmed errors are withdrawn and rounds 9 and 11 pass every criterion. False blocks were 0.9%-3.9% per round (2.3% overall). A third of the articles (38 of 120) drew on documents in 15 languages other than English; on them the desk performed as on English-only articles (false blocks 2.3% and 2.3%; planted errors not confirmed 98.5% and 98.9%), and of 221 sentences it confirmed on a passage not in English, 2

were factual errors under the publisher's standard. The remaining failures are mostly the model's: a fiscal year inside a credited figure, a source's own error repeated with its credit.

Conclusion. A deterministic-first desk reaches near-complete detection of planted errors on held-out articles at about one model call per three to four sentences checked, with every decision traceable to a document. Passing every criterion in a single round is not yet reliable: rounds differ by a few sentences, within the variation of the model between runs.

Keywords: automated fact-checking; local journalism; retrieval-augmented verification; language models; evaluation methodology; multilingual evidence

1. Introduction

Every article a newsroom publishes, whether a reporter or a language model drafted it, carries a specific risk: a sentence that says more, less or something other than the documents behind it. The risk is not abstract. A wrong date tells a reader to show up a week late; a quotation with one word changed puts words in an official's mouth; a figure credited to the wrong outlet misleads about who reported what. The editor who must catch these has the article and, at best, a folder of documents.

Automated fact-checking research has mostly framed the task as verifying a claim against a large reference corpus such as an encyclopedia (Thorne et al., 2018; Guo et al., 2022). A newsroom's task is narrower and stricter. The question is not whether a claim is true in the world, but whether the article states what its sources state: attribution fidelity. The evidence is known (the documents the newsroom held for that article), and the output is not a score but a decision an editor acts on: publish, fix, or look again.

This paper reports a system built for that task and its evaluation. The system's design principle is deterministic-first: code decides every sentence it can decide with certainty, and a model is asked only the questions code cannot answer, with only the passages code located, under a prompt that forbids background knowledge. Its evaluation principle is pre-registration on held-out material (articles kept out of development: never checked before their round, and not used to build any fix before it): each round's protocol, articles and criteria were committed to version control before the desk ran, the errors to be found were planted by an agent that never saw the desk, and the labels the desk was scored against were made by agents that never saw its verdicts.

Our contributions are (1) the architecture, deployed on a New York City publisher's newsroom service in October 2026; (2) a two-part evaluation protocol, planted errors and blind labelling of every sentence, applied in thirteen rounds; (3) the measured effect of 47 numbered fixes, each measured before it shipped or was set aside (F2, F4 and F35 were set aside); and (4) an account of what remains, which is mostly not code's to fix.

1.1 Existing fact checkers

Automated fact-checking began as support for professional fact-checkers judging public claims. ClaimBuster monitors interviews, speeches, debates, social media and news to identify factual claims and match them against a curated repository of professional fact-checks (Hassan et al., 2017); Nakov et al. (2021) survey the tools for each step of that work: finding claims worth checking, detecting claims already checked, retrieving evidence and verifying the claim. The benchmarks of this line ask whether a claim is true. LIAR holds 12.8K short statements labelled by PolitiFact (Wang, 2017); MultiFC, claims from 26 English fact-checking websites (Augenstein et al., 2019); FEVER, 185,445 claims made by altering Wikipedia sentences and verified against Wikipedia (Thorne et al., 2018); AVeriTeC, 4,568 real-world claims checked by 50 organizations, with evidence from the open web (Schlichtkrull et al., 2023); X-FACT, statements in 25 languages, on which the best model reached an F-score of about 40% (Gupta and Srikumar, 2021). Domain datasets carry the task into science (Wadden et al., 2020), public health (Kotonya and Toni, 2020) and finance (Rangapur et al., 2025). Guo et al. (2022) survey the field.

A second line checks text a model generated against the documents it was generated from. Abstractive summarizers were found to add content unfaithful to the input document (Maynez et al., 2020), and consistency checkers were built from rule-based transformations of source sentences (Kryscinski et al., 2020), sentence-level natural-language inference (Laban et al., 2022) and a single alignment function trained across tasks (Zha et al., 2023); TRUE compares such metrics across tasks (Honovich et al., 2022). For language-model output in general, AIS defines when a statement is attributable to identified sources (Rashkin et al., 2023); FActScore splits a generation into atomic facts and scores the share a knowledge source supports (Min et al., 2023); SAFE has a model check each fact with Google Search (Wei et al., 2024); RARR finds evidence with web search and revises unsupported content (Gao et al., 2023a); FacTool checks with external tools (Chern et al., 2023); SelfCheckGPT compares sampled responses without any external database (Manakul et al., 2023); and MiniCheck trains small models that check a claim against grounding documents at GPT-4's accuracy for a fraction of its cost (Tang et al., 2024). Retrieval-augmented generation has its own corpora and evaluators (Niu et al., 2024; Es et al., 2024; Gao et al., 2023b). Audits of deployed systems show the gap these methods address: in four generative search engines, 51.5% of generated sentences were fully supported by their citations (Liu et al., 2023).

The desk belongs to the second line and differs from it in four ways. Its evidence is closed and known, the documents the newsroom held for the article, with no web search, so a verdict says whether the article states what its sources state, not whether a claim is true in the world. Its output is a decision on each sentence that an editor acts on, with every excerpt the model cites verified against the document it names, not a score. Code settles a sentence before any model call where it can and re-checks dates, figures, negations and credits after the model reads. And it was

evaluated as deployed, on held-out articles, under protocols registered before each round. Pre-registered evaluations are uncommon in this literature; one is an audit of AI legal research tools (Magesh et al., 2025). Its evidence is also multilingual: X-FACT checks claims made in 25 languages, while here English sentences are checked against sources in fifteen.

2. System

2.1 Evidence

For each article, the desk assembles an evidence snapshot: every document the newsroom's evidence routes delivered for the article (the documents its writer was given among them), every document an editor attached, and every readable page the article links. Each document is held whole; the snapshot is content-addressed and stored. A document is labelled with its kind, class (a primary record, a subject's own page, another outlet's report, the newsroom's own archive), outlet, address and, where the stored record names one, its byline. The newsroom's own earlier articles are tier B: they may support retrieval but never confirm a sentence alone. The desk downloads nothing beyond the pages the article links.

2.2 Deterministic settlement

Every sentence of the headline, body, SEO title and SEO description is a unit. Code extracts its atoms (quotations, numbers, dates, times, names) and applies five rules, in the order a, b, e, c, d; the first that holds settles the sentence:

- **a. No checkable fact.** A question, or a sentence with at most one content word and no atom.
- **b. A quotation in no document.** The quotation is held verbatim, verbatim but for letter case, as in-order fragments of an elided quotation, or word for word with different punctuation; otherwise it is not held and the sentence blocks as a misquotation. A source's own bracketed insertion is not the speaker's words (F45).
- **c. A figure in no tier-A document.** A number, date or time no independent document holds, in a sentence whose other content one window of one tier-A document holds.
- **d. One passage holds the whole sentence.** Every content word within one window of one tier-A document, and every name, figure, year and count bound to the source sentence that holds the rest (F10, F13, F17, F18); never for a sentence that attributes, reorders names, drops a qualifier, or whose negations differ from its source's (F1, F36).
- **e. Nothing shares its content.** No held document shares the sentence's checkable content.

Rules a and d confirm; b and e block; c shows the sentence to the editor (minor) with the figure named. In the last five rounds, code confirmed 23-31 sentences per round without a model call, about one in eight confirmed sentences.

2.3 Model adjudication

The remaining sentences go to a model (claude-opus-4-8 in rounds 1-6, claude-opus-5-5 from round 7; effort low), in groups by paragraph, each with the passages code retrieved for it: up to six BM25 passages (Robertson and Zaragoza, 2009) widened to whole sentences, up to two extra passages from the unshown part of a high-ranked chunk that overlaps a passage already chosen (F37), up to two passages located by meaning in documents not in English, and short linked documents whole. The prompt asks for one result per sentence (supported, contradicted, not found, no checkable fact) with excerpts, the elements it could not find, and whether every missing element is a matter of wording. It states nine rules, among them: judge only against the passages; attribution is part of the claim; a quotation must be word for word, except a faithful translation; what follows directly from the passages (a weekday from a date, a borough from a neighbourhood) is stated.

Three readings are taken. Code verifies every cited excerpt against the whole document it names (a Han, kana or Hangul character counts twice toward the minimum length, F41), discards readings that fail, and maps the rest to a verdict and a severity. A sentence every reading finds stated only in other words is shown to the editor, not blocked (F28); two "not found" against one "supported" are shown (F33); two "supported" against one "contradicted" block. A sentence the model did not support is read again when code finds a passage it was not shown that holds what the readings named as missing (the second look).

2.4 Checks after adjudication

Every confirmed sentence is checked again by code: its relative dates against its documents' dates; relative time words ("this weekend", "on Monday") that name another period when the article is read than when it was written, unless the article dates that weekday elsewhere, ties it to a dated event, or states a weekly schedule (F20, F26, F30, F34, F38, F40, F42, F44, F46, F47); the end of a stated date range; a "no date announced" a document contradicts; every outlet it credits against the documents from that outlet; and a duration ("nearly a century ago") against a year the article gives the same event (F43, shown, never blocking).

2.5 Evidence in other languages

Sixty-six of 208 articles checked in the week before the evaluation held at least one document not in English. In round 1, eight of nine false blocks fell in the two articles sourced in Spanish or Italian, because BM25 matches English words. The desk now locates passages in documents not in English by meaning with a local multilingual embedding model (Wang et al., 2024), loaded from disk without download (F6); shows short non-English documents whole (F21); matches quoted titles in inflecting languages on word stems (F31); and accepts a faithful translation of a quotation held in its original.

2.6 Recording, replay and cost

Every model reading is stored under a key of everything the model was shown (prompt version, model, effort, the sentence, its paragraph, each passage and its document label) and replayed on later runs instead of asking again. A production re-check of an unchanged article costs few or no model calls. A first check of a ten-article round cost 61-83 calls and 1.07-1.63 million input tokens in rounds 8-13.

2.7 What blocks

An open claim whose verdict is misattributed, inaccurate, unsupported, misquoted or needs-review at severity major or critical blocks publication; minor findings are shown to the editor. The editor sees the sentence, the verdict, the documents and excerpts that decided it, and a suggested fix.

3. Evaluation method

3.1 The standard

The desk checks attribution fidelity: a sentence is supported when its held documents state it. The labels it is scored against add a second question, asked of every sentence not supported: does it tell the reader something wrong? The rubric counts as a factual error a wrong number, date, time, name, title, place or count; a statement credited to a person or outlet that did not make or report it; quoted words that change what the speaker said; a claim the evidence contradicts; and a sentence that contradicts its own article. It does not count wording or paraphrase that keeps the facts, another name form of the right outlet, a fact that follows from stated facts, an evaluative characterization, or a general word that covers the actual fact. Materiality is minor, moderate (a wrong detail a reader could act on or repeat) or major (it changes the news).

The publisher, who owns the standard, ruled on its boundaries during the study: misciting counts as a factual error for this evaluation; omissions, boundary calls and subject nuances do not (from round 6); hair-splitting does not ("Puccini and workshop plural are hair splitting"); and stale verb tense is not a factual error. A relative time word that names another day when the article is read is a date, not a tense, and still counts.

3.2 Rounds

Each round drew ten articles at random (a seeded sample, seed in the protocol) from the publisher's queue of articles awaiting review that had never been fact-checked, and committed its protocol (the system under test by commit and policy version, the articles, the criteria, the reference date on which relative times are read) before any run. Of the 130 articles, 127 were drafted by the publisher's language-model writer; the other three, published earlier, carry no drafting

record. The desk ran once on the production service (Part B), then in process with each article changed in memory and rolled back (Part A). Nothing was published, and no run on other code used the desk while a round ran.

Part A: planted errors. A writer agent that never saw the desk, given only the evidence and a draw order of sentences, walked the order and changed sentences into false ones under fixed quotas: eight wrong numbers (N), eight wrong dates or times (D), eight wrong people (P), six wrong places or organizations (L), six reversed relations (R), four wrong attributions (A), and, where a quotation is held only in another language, a mistranslated quotation (T). Each change is one sentence, reproducible by string replacement, with the evidence for the original quoted verbatim and a search of the evidence showing the changed fact absent. Ten more sentences were rewritten without changing a fact. Criteria: A1, at least 36 of 40 planted errors not confirmed; A2, at least 32 blocked; A3, every rewrite whose original was supported confirmed (all but one allowed).

Part B: every sentence. Four reader agents, two per half of the articles, labelled every sentence of the registered run against the same evidence the desk held, without its verdicts: supported, partly, not in evidence, contradicted, or no fact, with every element, the document and a verbatim quotation for each, and for every sentence not supported the factual-error label and materiality. Disagreements and every sentence not supported by both were adjudicated against the rubric and earlier rulings, with checks outside the evidence where the evidence could not settle a fact (a byline, a deadline, a vote count), recorded with raw copies. Criteria: B1, no factual error confirmed; B2, at most 2% of all sentences correct but blocked. Proportions are reported with 95% Wilson score intervals (Wilson, 1927), per round from the round's own counts and over rounds 2-13 from the summed counts; the intervals treat sentences as independent, and sentences of one article are not, so they understate the uncertainty. Reader agreement is Cohen's kappa (Cohen, 1960).

3.3 Development between rounds

Each round's failures were traced to their mechanism, and each fix was measured before it shipped: offline over the registered runs of every earlier round and a population of 218 production runs (deterministic changes, every changed decision listed with its answer-key label), and, for changes to what the model is shown, by running the desk in memory on the answer keys of every earlier round and comparing sentence by sentence. A fix shipped only when no confirmed error was added; F35 (the credited source's own passages for the model) was measured and not shipped. The next round then measured the deployed policy on articles none of the fixes had seen.

3.4 Agents and the adjudicator

The writer and the readers are language-model agents (Anthropic Claude), each started fresh for one task with a written brief and the rule to read only the evidence and its own inputs; their briefs and reports are in the package. The adjudicator is the language-model agent that ran the study; in rounds 10, 12 and 13 no desk verdict was read before the adjudication was written, and where it was (round 11, ten sentences), the rulings applied rules already written and the disclosure is in the round's deviations. The publisher's rulings are quoted verbatim in the package. No human labelled sentences.

4. Results

4.1 Every round

Table 1 gives each round as registered and, for B1, after the publisher's rulings. The reclassification does not edit the registered answer keys; it is recorded beside them with a reason for each sentence.

Table 1. Test rounds 2-13 (registered results; B1 also after the publisher's rulings of 28 October 2026). A1: planted errors not confirmed; A2: planted errors blocked; A3: rewrites confirmed, of the rewrites whose original sentence was adjudicated supported (every rewrite in round 2); B1: factual errors the desk confirmed; B2: correct sentences blocked (share of the round's sentences, with its Wilson 95% interval). Model: claude-opus-4-8 or claude-opus-5-5, effort low.

Round	Policy	Model	Sentences	A1	A2	A3	B1 registered	B1 after rulings	B2 (95% interval)	Errors in the copy caught
2	v4p14	opus-4-8	306	40/40	39/40	10/10	1	1	10 (3.3%; 1.8-5.9%)	0 of 1
3	v4p16	opus-4-8	251	38/40	38/40	10/10	5	4	4 (1.6%; 0.6-4.0%)	0 of 5
4	v4p18	opus-4-8	214	39/40	39/40	8/9	3	1	2 (0.9%; 0.3-3.3%)	2 of 5
5	v4p19	opus-4-8	236	38/40	38/40	10/10	3	1	5 (2.1%; 0.9-4.9%)	0 of 3
6	v4p20	opus-4-8	173	40/40	40/40	8/9	6	4	5 (2.9%; 1.2-6.6%)	1 of 7
7	v4p24	opus-5-5	229	40/40	40/40	7/7	4	2	9 (3.9%; 2.1-7.3%)	4 of 8
8	v4p26	opus-5-5	281	41/41	41/41	10/10	2	0	8 (2.8%; 1.5-5.5%)	11 of 13
9	v4p28	opus-5-5	248	39/40	39/40	10/10	1	**0**	**3 (1.2%; 0.4-3.5%)**	2 of 3
10	v4p29	opus-5-5	245	40/40	40/40	9/9	**0**	0	8 (3.3%; 1.7-6.3%)	4 of 4
11	v4p31	opus-5-5	263	40/40	40/40	10/10	1	**0**	**5 (1.9%; 0.8-4.4%)**	7 of 8
12	v4p32	opus-5-5	269	41/41	41/41	9/9	2	1	6 (2.2%; 1.0-4.8%)	1 of 3
13	v4p33	opus-5-5	219	40/40	40/40	10/10	2	1	3 (1.4%; 0.5-4.0%)	12 of 14

Round 1 (v4p9) used earlier definitions and no planted errors (B1 as confirmed sentences labelled contradicted, 5 of 225; B2 as confirmed sentences only partly supported, 9.3%); it is reported in the package and excluded from the totals below.

Rounds 9 and 11 pass every criterion under the publisher's standard: round 9 with no confirmed factual error, 3 false blocks of 248 and 39 of 40 planted errors blocked; round 11 with no confirmed factual error, 5 of 263 and 40 of 40. Registered, round 10 was the only round with no confirmed factual error.

4.2 Planted errors

Over rounds 2-13 the desk did not confirm 476 of 482 planted errors (98.8%; 95% interval 97.3-99.4%) and blocked 475 (98.5%; 97.0-99.3%). The six it confirmed were each traced to a code rule and fixed: a name bound to another sentence's description (fixed by F10), a bare year never compared (F13), a date held in a list beside the right one (F17), a count written in words (F18), a moved negation (F36), and one reversed idiom read once (all readings since v4p26). No class recurred, and the last four rounds confirmed none. The rewrites were confirmed in 111 of 113 cases (98.2%; 93.8-99.5%).

4.3 Confirmed factual errors

The 30 registered confirmed errors fall into four groups. Fifteen are withdrawn under the publisher's rulings: three by direct ruling (a double bill described as two operas created for an anniversary; a premiere program's title; "workshops" plural), one as stale tense, and eleven by the adjudicator on the same standard (a neighbourhood boundary, per-count wording of a stated penalty, a program described by what its operator serves, an omission, a summary's compression, a dropped "only", a stale "this weekend" beside the correct date, a paraphrase of the source's own characterization, a ballot measure described by its official title, an approximation the primary source uses, a direction on which two sources disagree). Of the 15 kept, four are relative time words with no date carried from a source written days or months earlier (caught by code since F20 and its successors, the last of them, "on Friday and again on Sunday", by F46); four are misciting (a statement credited to the outlet that republished it, or a detail credited to the one of two outlets that did not report it); seven are wrong statements: an instruction to buy insurance from a service that sells none, a logo said to be still unreleased, a term count, a legal status before signature, a year frame, an outdated credential, a figure credited to a report for the wrong fiscal year.

Re-run in memory on the answer keys of rounds 2-10, the policy then deployed (v4p30) still confirmed 4 of the 25 registered errors of those rounds; all four were later withdrawn by the rulings.

4.4 False blocks

False blocks were 68 of 2,934 sentences over rounds 2-13 (2.3%; 95% interval 1.8-2.9%), between 0.9% and 3.9% per round. Their causes changed as they were fixed. Code defects found and fixed: a stale-time check that read schedules and recurring weekdays as dates (F42, F44, F47), a passage that ended where the supporting line began (F37: four of round 10's eight false blocks), a 12-character excerpt minimum calibrated for letters that discarded verbatim Chinese quotations (F41), a source's bracketed insertion treated as the speaker's words (F45), and a byline the capture held but never served (F39: thirteen true sentences naming a column's author blocked in one article). What remains is mostly the model's strictness: a plain-language para-

phrase, a service line ("documents are posted on the agency's site"), a figure a reader derives by subtraction.

4.5 Code and model

In rounds 9-13 code's settlement rules confirmed 23-31 sentences per round with no model call and caused one false block (round 12: a quotation without the source's bracketed insertion, fixed by F45); the checks after adjudication caused one in each of rounds 10-13 (a weekday the article dates by arithmetic, a recurring schedule twice, a generic weekday), each fixed. In rounds 10-13 every planted error was blocked by the model's readings; code's role in Part A is to block what the model would wave through (the moved negation of round 9, the date in a list of round 5) and to bind what the model sees.

4.6 Variation between runs

Each round runs Part B once and Part A once on the same code. Co-paragraph control sentences, unchanged but sharing a paragraph with a planted change, changed outcome between the two runs in 2 to 8 of 63 to 91 per round (rounds 6-13). A round's B2 can therefore move by a few sentences between runs of the same code, which is the same order as the margin by which rounds 10 and 12 failed B2.

4.7 Cost

A ten-article round's first check took 61-83 model calls, 1.07-1.63 million input tokens and 174,000-269,000 output tokens in rounds 8-13, about one call per three to four sentences checked and one per seven sentences confirmed. The production desk replays recorded readings; a re-check of an unchanged article makes few or no calls.

4.8 Evidence in other languages

A third of the articles drew on documents not in English: 38 of the 120 articles of rounds 2-13 held at least one, 93 documents in 15 languages (Spanish 39, Chinese 17, Urdu 7, Polish 6, Korean 5, Bengali 3, Arabic 3, Russian 3, Italian 3, Haitian Creole 2, Japanese 1, Greek 1, Albanian 1, French 1, Hindi 1). A document counts as not in English when the desk's own rule (F6) says so and a language identifier (lingua) does not read it as English; three corrections made by reading the documents (two in Haitian Creole that the identifier named otherwise, one language-selection menu) are listed in the analysis script. Table 2 splits the registered results of rounds 2-13 by article.

Table 2. Rounds 2-13 by whether the article held a document not in English (counts with 95% Wilson intervals). A1, A2: planted errors not confirmed, blocked; A3: rewrites confirmed (as in Table 1); B2: correct sentences blocked; B1 after rulings: factual errors confirmed, of the sentences the desk confirmed.

Articles	n	Sentences	A1	A2	A3	B2	B1 registered	B1 after rulings
With a document not in English	38	1,194	202/205 (98.5%; 95.8-99.5%)	201/205 (98.0%; 95.1-99.2%)	46/48	28 (2.3%; 1.6-3.4%)	10	5 of 1,070 (0.5%; 0.2-1.1%)
English only	82	1,740	274/277 (98.9%; 96.9-99.6%)	274/277 (98.9%; 96.9-99.6%)	65/65	40 (2.3%; 1.7-3.1%)	20	10 of 1,493 (0.7%; 0.4-1.2%)

The two groups do not differ beyond their intervals. At the sentence level, the desk confirmed 221 sentences on a passage of a document not in English, 200 of them on such documents alone; 4 were registered as factual errors and 2 remain under the publisher's standard (0.9%; 95% interval 0.25-3.2%). Neither is a misreading of the other language: the model read the Urdu and the Spanish correctly and accepted a word that changes a fact (a second term where the source says re-election; a bill "made the law" where the source says it awaits the mayor's signature). Both are from round 6, and both are among the examples of prompt rule 7, added after that round. Round 1, before passages by meaning (F6) and whole short documents (F21), had eight of its nine false blocks in the two articles sourced in Spanish or Italian.

5. Discussion

What the deterministic-first design buys. Code decides with certainty what it can and makes the model's task narrow: one sentence, the passages located for it, nine rules. Every verdict is traceable to a document and an excerpt code has verified. Code also catches what a model reading a single sentence will not: arithmetic on dates, a negation that moved, a figure written in words, a weekday a week stale. The planted-error results (98.8% not confirmed over twelve held-out rounds) show the combination does not wave through errors of the kinds a writer makes.

What it does not buy. Under attribution fidelity, a source's own error repeated with its credit is supported: an outlet's wrong vote shares, a presenter's approximate anniversary. The desk confirms them because its documents state them. We recorded such cases but did not count them, because the protocol asks whether the article states what its sources state. A desk that checks world truth needs evidence the newsroom did not hold.

Most remaining failures are the model's reading. A fiscal year inside a credited figure, a scope qualifier the model was shown and read past, a gloss it wanted defined: these are judgments of meaning, and the deterministic layer does not reach them without rules that would

block correct sentences. The model's verdicts also vary between runs on the same passages; all three readings are taken for every sentence the model judges, and a split reading is shown to the editor, not blocked.

Evidence capture is part of fact-checking. Two of the most consequential failures were in the evidence, not the desk: a captured page that dropped its byline (thirteen true sentences blocked in one article) and a page whose visible byline named the photographer while its metadata named the writer (true attributions blocked with a fix that would have re-attributed the piece). The first is fixed by serving the byline the stored record names (F39); the second is open.

The standard is the publisher's. The rubric changed during the study by the publisher's rulings: misciting counts; omissions, hair-splitting and stale tense do not. Rulings made after results are known risk fitting the standard to the outcome. We report the registered results unedited beside the reclassified ones, list every withdrawn sentence with its reason, and note that the rulings that decide passes (hair-splitting, stale tense) also withdrew errors from rounds that still fail. The desk's own prompt still instructs the model to treat two of the withdrawn kinds (a dropped "only"; "last seen in" for "reminiscent of") as unstated. A prompt aligned with the ruling (v4p35) was measured on the answer keys of rounds 2-13 against the policy before it: false blocks fell from 34 to 32 of 2,934 sentences, and confirmed errors rose from 2 to 3. It let through two errors the deployed prompt blocks (a newsletter sent to many readers credited as an interview given to one outlet, and a stale "the previous week") and blocked one the deployed prompt lets through. It was not deployed: accuracy comes before the false-block rate.

Beyond journalism. The desk's question, whether a drafted text states what its source documents state, is not specific to news. Wherever a language model drafts from a known set of documents and an error has a cost, the same design applies: a closed evidence set; code that settles quotations, figures and dates; a model that reads only located passages; excerpts verified against the documents they name; and a recorded, replayable decision for every sentence. We have not tested the desk outside the newsroom. What follows is where it applies and what would have to change.

Finance. Analysts' notes, earnings summaries, client reports and filings restate figures from documents their writer holds. The errors the desk is built to catch, a figure, period or attribution that differs from its source, are the costly ones there: on questions about public companies' filings, GPT-4-Turbo with a retrieval system answered wrongly or refused 81% of the time (Islam et al., 2023), and Fin-Fact pairs financial claims with professional fact-checkers' verdicts (Rangapur et al., 2025). The deterministic rules transfer most directly: a quotation or figure no document holds, a fiscal period stated as another (one of the errors this desk kept was a figure credited to a report for the wrong fiscal year), a relative date. New rules would be needed for units and scale, for figures derived by arithmetic (growth rates, margins) that code can recompute, and for tables, which the desk reads as text.

Law. Briefs and memoranda cite authorities their writer can hold. Public language models hallucinate on specific questions about federal court cases at least 58% of the time (Dahl et al., 2024), and in a pre-registered evaluation, legal research tools built on retrieval hallucinated between 17% and 33% of the time (Magesh et al., 2025). A desk holding the cited authorities would check each quotation word for word and each stated holding against the passage cited; it could not judge whether an authority is still good law, which needs evidence beyond the documents held.

Medicine, compliance and audit. Clinical summaries drawn from a patient's record, regulatory submissions drawn from study reports, and audit or compliance reports drawn from source records share the structure: a known evidence set, a drafted text, and a cost when the two differ. Med-HALT tests language models for hallucination on medical examination questions from several countries (Pal et al., 2023). Two features of the desk matter more in these fields than in news: it reads sources in other languages by meaning, and it keeps every decision with the passages behind it, a record a reviewer or regulator can replay. What does not transfer is the publisher's rubric. What counts as an error (here, stale tense does not) is set by each field's own standard, and the evaluation of this paper, with planted errors and blind labels, would have to be run again against it.

Agents that answer for a company. The same failure appears when an agent speaks for a business. An airline's website chatbot told a customer that a bereavement fare could be claimed retroactively, which the airline did not permit; the tribunal held the airline responsible, finding that it "did not take reasonable care to ensure its chatbot was accurate" (*Moffatt v. Air Canada*, 2024). A software company's AI support agent told a user that a subscription allowed one device "as a core security feature", a policy the company did not have (Edwards, 2025). Agents are moving into this work quickly: 40% of respondents from organizations with revenue over $1 billion report scaling AI agents (McKinsey & Company, 2026), and Gartner predicts that a third of enterprise software applications will include agentic AI by 2028, while more than 40% of agentic AI projects will be canceled by the end of 2027 for escalating costs, unclear business value or inadequate risk controls (Gartner, 2025). A company's policies, prices, contracts and procedures are a closed evidence set like a newsroom's folder. An agent's answer can be checked against them before it is sent: a statement no document supports is the case this desk blocks, and the decision is kept with the passage that settled it.

6. Limitations

The desk reads only an article and its documents, so nothing in it depends on who wrote the article, but the test articles were drafted almost entirely by one language-model writer (127 of 130). Its accuracy on reporters' copy, whose errors may differ in kind and rate, is untested. All labelling was done by language-model agents, not human annotators; their agreement (kappa

0.74-0.96 per round) is between agents of one model family, and the adjudicator is an agent of that family too. The publisher's rulings, not an external standard, define a factual error. Fixes were chosen on each round's failures and measured on the next round's articles; every round is held out from the fixes it tests, but the policy as a whole was developed on the same publisher's articles, and the evaluation is of one newsroom's New York City local news. Each round is ten articles; rounds differ by a few sentences, and the B1 and B2 intervals are wide: a round with no confirmed factual error still has a 95% upper bound of 1.8% on the share of its confirmed sentences that are wrong (round 10: 0 of the 206 it confirmed), and a round's B2 interval spans three to five and a half percentage points (Table 1). Part A's errors are written by an agent and may be easier or harder than a writer's own. The pool of never-checked articles was exhausted after round 13, so the latest policy (v4p36) has been measured only offline and in production checks. Table 2 splits by article, not by the evidence each sentence rests on; its language names come from an identifier corrected by reading only where it was implausible, and most of the fifteen languages have one to seven documents, too few to measure any one of them.

7. Ethics and data availability

The articles were drafts in the publisher's review queue, written from the newsroom's sources: 127 of the 130 had never been published, and three had been published earlier (February to May 2026) and were back in the queue when checked. The evidence is what the newsroom held: press releases and newsletters it received by e-mail (64 documents in the thirteen rounds), agency pages, public records and other outlets' published reports. The publisher regards this material as public information; it is nonetheless mostly third-party work, and none of it is reproduced here. Cases are described in broad strokes, without the names of private individuals. A de-identified dataset (sentence-level labels, desk verdicts, planted changes and per-round criteria) is released with this paper: the names of private individuals are replaced by per-article placeholders, and residential addresses, e-mail addresses and telephone numbers are removed; public officials and public figures named or quoted for their public role are kept. The full research package (every protocol, run, reading, answer key, adjudication, measurement script and the code at each policy, with a SHA-256 manifest and the chain of evidence for every number) holds the evidence documents and is available to reviewers on request. Language-model agents were used to build the system, to plant errors, to label sentences and to draft this manuscript; the author is responsible for its content.

8. Conclusion

A fact-check desk that settles what it can with code and asks a model only the rest reached near-complete detection of planted errors on twelve rounds of held-out local news articles, cut confirmed factual errors from five or six per round to zero to two, and kept false blocks near 2%,

with every verdict traceable to a document. Two rounds passed every criterion under the publisher's standard. The remaining failures are mostly the model's reading of meaning and the evidence's own errors, which is where further work should go: better capture, and a model that is told what it most often misreads.

References

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., and Simonsen, J. G. (2019). MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. *Proceedings of EMNLP-IJCNLP 2019*, 4684-4696.

Chern, I.-C., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., He, J., Neubig, G., and Liu, P. (2023). FacTool: factuality detection in generative AI -- a tool augmented framework for multi-task and multi-domain scenarios. arXiv:2307.13528.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. (2024). Large legal fictions: profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64-93.

Edwards, B. (2025, April 17). Company apologizes after AI support agent invents policy that causes user uproar. *Ars Technica*.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). Ragas: automated evaluation of retrieval augmented generation. *Proceedings of EACL 2024: System Demonstrations*, 150-158.

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D.-C., and Guu, K. (2023a). RARR: researching and revising what language models say, using language models. *Proceedings of ACL 2023 (Long Papers)*, 16477-16508.

Gao, T., Yen, H., Yu, J., and Chen, D. (2023b). Enabling large language models to generate text with citations. *Proceedings of EMNLP 2023*, 6465-6488.

Gartner (2025, June 25). Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027 [Press release]. Gartner, Inc.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10, 178-206.

Gupta, A., and Srikumar, V. (2021). X-FACT: a new benchmark dataset for multilingual fact checking. *Proceedings of ACL-IJCNLP 2021 (Short Papers)*, 675-682.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., and Tremayne, M. (2017). ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12), 1945-1948.

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., and Matias, Y. (2022). TRUE: re-evaluating factual consistency evaluation. *Proceedings*

of NAACL 2022, 3905-3920.

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., and Vidgen, B. (2023). FinanceBench: a new benchmark for financial question answering. arXiv:2311.11944.

Kotonya, N., and Toni, F. (2020). Explainable automated fact-checking for public health claims. *Proceedings of EMNLP 2020*, 7740-7754.

Kryscinski, W., McCann, B., Xiong, C., and Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. *Proceedings of EMNLP 2020*, 9332-9346.

Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A. (2022). SummaC: re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10, 163-177.

Liu, N., Zhang, T., and Liang, P. (2023). Evaluating verifiability in generative search engines. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7001-7025.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(2), 216-242.

Manakul, P., Liusie, A., and Gales, M. (2023). SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. *Proceedings of EMNLP 2023*, 9004-9017.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of ACL 2020*, 1906-1919.

McKinsey & Company (2026, August 25). The state of AI in 2026: on the road to ROI [Survey report].

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). FActScore: fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of EMNLP 2023*, 12076-12100.

Moffatt v. Air Canada, 2024 BCCRT 149 (Civil Resolution Tribunal of British Columbia).

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., and Da San Martino, G. (2021). Automated fact-checking for assisting human fact-checkers. *Proceedings of IJCAI 2021*, 4551-4558.

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., and Zhang, T. (2024). RAGTruth: a hallucination corpus for developing trustworthy retrieval-augmented language models. *Proceedings of ACL 2024 (Long Papers)*, 10862-10878.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. (2023). Med-HALT: medical domain hallucination test for large language models. *Proceedings of CoNLL 2023*, 314-334.

Rangapur, A., Wang, H., Jian, L., and Shu, K. (2025). Fin-Fact: a benchmark dataset for multimodal financial fact-checking and explanation generation. *Companion Proceedings of the ACM Web Conference 2025*, 785-788.

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., and Reitter, D. (2023). Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4), 777-840.

Robertson, S., and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333-389.

Schlichtkrull, M., Guo, Z., and Vlachos, A. (2023). AVeriTeC: a dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks)*, 65128-65167.

Tang, L., Laban, P., and Durrett, G. (2024). MiniCheck: efficient fact-checking of LLMs on grounding documents. *Proceedings of EMNLP 2024*, 8818-8847.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. *Proceedings of NAACL-HLT 2018*, 809-819.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or fiction: verifying scientific claims. *Proceedings of EMNLP 2020*, 7534-7550.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual E5 text embeddings: a technical report. arXiv:2402.05672.

Wang, W. Y. (2017). "Liar, liar pants on fire": a new benchmark dataset for fake news detection. *Proceedings of ACL 2017 (Short Papers)*, 422-426.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., and Le, Q. (2024). Long-form factuality in large language models. *Advances in Neural Information Processing Systems 37*, 80756-80827.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209-212.

Zha, Y., Yang, Y., Li, R., and Hu, Z. (2023). AlignScore: evaluating factual consistency with a unified alignment function. *Proceedings of ACL 2023 (Long Papers)*, 11328-11348.