

FactRot: Auditing Typed Fact Survival Under Iterative LLM Rewriting

Vladislav Kulikov^{1,2}, Mikhail Mozikov³, Aleksei Stepin², Boris Tolstokulakov²,
Artyom Kabanov^{1,2}, Ivan Nasonov⁴, Nikita Glazkov³, Ilya Makarov^{1,3}

¹Research Center of the Artificial Intelligence Institute, Innopolis University, Innopolis, Russia

²AI Talent Hub, ITMO University, Saint Petersburg, Russia ³AXXX, Moscow, Russia

⁴MIRIAI, Moscow, Russia

Correspondence: vkulikov@niuitmo.ru

Abstract

FactRot measures whether typed factual mentions survive iterative LLM rewriting. We follow each ledger fact through every hop of a rewrite chain and score it separately, instead of measuring aggregate drift in the rewritten text. We score 1,113 ledger facts from 25 Wikipedia passages across 4 models, 4 rewrite operations, and 3 prompt conditions, with chains reaching depth 12. A GEE with passage clusters finds per-hop decay (OR=0.743 per additional hop, $p < .001$). Type-specific faithfulness instructions do not measurably outperform generic faithfulness at hop 5 ($\Delta = 0.026$, 95% CI [-0.022, 0.073]); both improve survival over neutral prompting by roughly 0.13–0.16 absolute. On the depth-12 neutral subset, answer-restate rewriting is much more destructive than direct paraphrasing, and the frontier preview model has the highest hop-12 survival among the four models. Exact fact-type rankings, however, are unstable across models (Kendall $W = 0.082$ at hop 5).

1 Introduction

LLM outputs are often rewritten repeatedly: summaries are expanded, reports are restated, retrieved evidence is compressed and passed downstream. In these chains, a small factual drift at one hop becomes the input to the next. Prior “telephone game” studies show that iterative generation changes aggregate textual properties and distorts transmitted information (Perez et al., 2025; Mohamed et al., 2025; Ghafouri and Ferrara, 2026). For many applications the operational question is narrower: whether a given number, date, entity, negation, conditional, or causal claim survived.

FactRot studies this question with a typed fact ledger and per-hop survival scores over a cleaned panel of 339,320 measurements covering 1,113 facts from 25 passages, 4 models, 4 rewrite operations, and 3 prompt conditions, with chains reaching depth 12. Hop-0 survival is 0.995, and 1.01% of

the rows in that panel are unparseable judge outputs. Our contributions are threefold. First, we introduce a typed fact-survival protocol for iterative rewriting, linking each extracted fact to a per-hop survival trajectory. Second, we show that short-chain faithfulness prompting substantially improves survival over neutral rewriting, while enumerating fact types adds no measurable gain over a generic faithfulness instruction at hop 5 ($\Delta = 0.026$, 95% CI [-0.022, 0.073]). Third, we quantify how rewrite intent and model choice affect deep-chain survival, with answer-restate rewrites losing substantially more facts than direct paraphrase on the neutral depth-12 subset.

2 Related Work

Iterative-rewriting work treats repeated LLM generation as a telephone game. Perez et al. (2025) show that multi-turn transmission chains drift toward cultural attractors in aggregate properties of the generated text; Mohamed et al. (2025) measure how distortion accumulates over iterative translation chains; and Ghafouri and Ferrara (2026) study how social information is transmitted and lost in model-to-model communication. FactRot changes the unit of analysis: it follows the same typed ledger fact through every rewrite hop instead of summarizing each generated text as a whole. This yields per-fact survival trajectories that can be compared across rewrite operations, models, prompt conditions, and depths.

Fine-grained factuality metrics such as FactScore decompose generated text into atomic claims for support checking (Min et al., 2023), and related long-form work scales claim-level factuality scoring with LongFact and the SAFE evaluator (Wei et al., 2024); FACTOR converts a reference corpus into a contrastive factuality benchmark (Muhlgay et al., 2024); Chain-of-Verification reduces hallucinations through model-internal

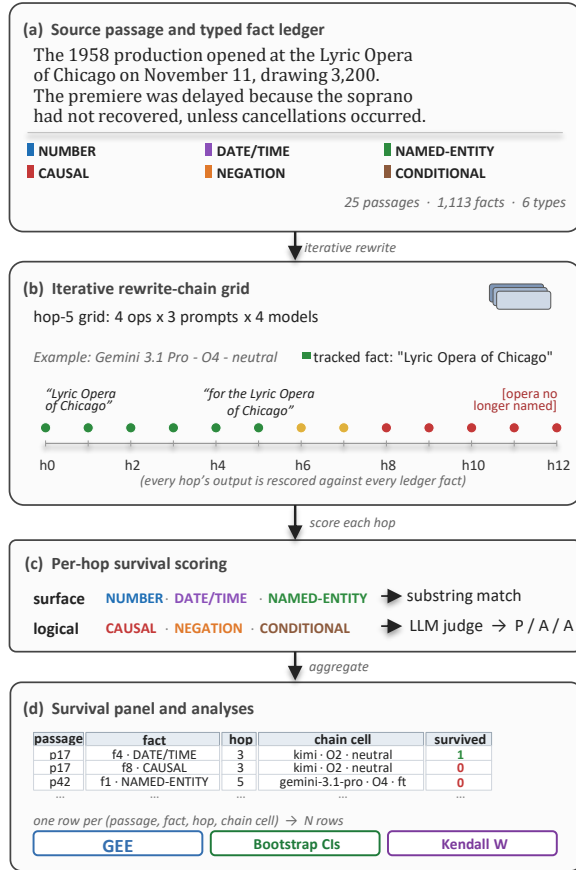


Figure 1: **The FACTROT framework.** (a) A Wikipedia passage is paired with a typed fact ledger covering six categories. (b) Each passage is rewritten through a 4-op $\times$ 3-prompt $\times$ 4-model design grid; the strip tracks one NAMED-ENTITY mention across 13 hops of a single answer-rewrite chain. (c) Surface facts are scored by normalized substring matching, logical facts by an LLM judge (*present* / *altered* / *absent*). (d) The resulting per-(passage, fact, hop, chain cell) panel feeds passage-clustered GEE, cluster-bootstrap CIs, and Kendall’s W .

recheck prompts (Dhuliawala et al., 2024). Other work probes factual knowledge through relation-level queries and studies consistency across paraphrases of those queries (Petroni et al., 2019; Elazar et al., 2021). FactRot instead follows a fixed ledger of typed facts longitudinally, re-scoring each fact after every rewrite hop.

3 Method

Design overview. Figure 1 summarizes the pipeline. We first build a typed ledger for each source passage, then run iterative rewrite chains that vary the model, rewrite operation, prompt condition, and depth. The same ledger facts are scored against every hop output, producing a survival panel whose rows identify a passage, fact,

hop, model, operation, and prompt condition. This design separates short-chain prompt effects at hop 5 from the neutral deep-chain subset used for hop-12 operation and model contrasts.

Corpus and facts. Passages come from the English Wikipedia snapshot of 2023-11-01 (Wikimedia Foundation, 2023). Articles are streamed in shuffled order with a fixed seed (42), up to 15,000 articles, skipping disambiguation pages, titles beginning “List of”, and articles shorter than 300 words. A passage is a contiguous sentence-aligned span of 120–180 words, and candidates yielding fewer than eight extracted facts are discarded. The surviving candidates are ordered deterministically by fact-type coverage and then by fact count, and the first 300 form the study corpus; all of them contain all six fact types. The hop-5 panel uses the first 25 passages of this ordering, and the deep chains use the first 8.

Each passage is paired with a typed ledger covering six categories: NUMBER, DATE/TIME, NAMED-ENTITY, CAUSAL, CONDITIONAL, and NEGATION. Surface facts are span-level mentions extracted with spaCy en_core_web_sm named-entity recognition under fixed label mappings (for example, CARDINAL, PERCENT, MONEY and QUANTITY become NUMBER), supplemented by two regular expressions that recover numeric and date mentions the tagger misses; each mention is stored with a normalized canonical value. Logical facts are sentence-level mentions selected by trigger lists, so the lists also fire on sentences that only superficially express the relation. Appendix C states the label mappings, regular expressions, normalization rules, and trigger lists in full.

Rewrite chains. At each hop the previous text is fed back through one of four rewrite operations. **O1** asks for a faithful paraphrase, keeping length and structure roughly intact. **O2** compresses the passage into a short summary and then re-expands the summary into a full passage. **O3** rewrites the passage as a third-party report. **O4** synthesizes an implicit question from the passage and produces an answer to it; this operation therefore rebuilds the surface form from extracted content, and we expected peripheral mentions to suffer most under it. Because each operation is conveyed by a different instruction phrase (for example, O1 contains “paraphrase faithfully” while O4 contains “answer then restate”), the operation effect we report is the joint effect of the rewrite intent and the wording

that conveys it; we cannot separate the two within this panel.

The four models cover two tiers. The flash tier contains three instances, deepseek-v4-flash, gemini-3-flash-preview and kimi-k2.6; the frontier tier is represented by one preview model, gemini-3.1-pro-preview.

The 3 prompt conditions differ only in a single instruction line. *Neutral* asks the model for the rewrite and nothing else. *Generic faithful* adds one sentence telling the model to preserve every factual claim in the input. *Fact-type-aware* replaces that sentence with one that enumerates the six categories and asks the model to preserve members of each by name. All three conditions are scored at hop 5; the neutral deep-chain subset reaches hop 12 on 8 passages and is reported separately.

Scoring. At each hop, a fact survives if its canonical surface form is found by a normalized substring matcher (NUMBER, DATE/TIME, NAMED-ENTITY) or an LLM judge returns PRESENT (CAUSAL, CONDITIONAL, NEGATION). Judge ALTERED/ABSENT verdicts score 0; unparseable outputs are missing data. Hop 0 is the untouched source passage and serves as the no-degradation baseline.

We perform three measurement-integrity checks: hop 0 remains near ceiling (0.995), unparseable judge outputs are treated as missing rather than as losses (1.01% of the clean panel), and every analysis uses the panel restricted to (passage_id, fact_uid) pairs present in the canonical ledger.

Judge and automated reliability. The judge scores only the three logical types; NUMBER, DATE/TIME, and NAMED-ENTITY never reach it. The primary judge is MiniMax M2.7 (minimax/minimax-m2.7), called at temperature 0 on batches of at most five facts. Two automated checks accompany it. A judge from a different family, GLM-5.1 (z-ai/glm-5.1), rescored a random subsample and matched the primary judge’s binary survival verdict (agreement 0.883, $\kappa = 0.62$, $n = 197$); rerunning the primary judge at temperatures 0.0 and 0.7 reproduced its own verdict (agreement 0.95). Both measure agreement between automated raters, not correctness.

To validate the judge against human annotations, 3 authors independently labeled 200 uniformly sampled logical-fact cells by hand, covering 88 distinct facts across all 25 passages, using the same PRESENT/ALTERED/ABSENT rubric and blinded to the judge verdict, rewrite model, opera-

tion, and prompt condition. On the binary survival outcome used in every downstream analysis, the judge matches the majority-vote human label with agreement 0.865 ($\kappa = 0.516$, passage-clustered bootstrap 95% CI [0.334, 0.658]). Pairwise agreement among the annotators themselves ranges from $\kappa = 0.422$ to 0.670 on the same cells, which places judge-human agreement inside the range the annotators span. Most residual disagreement sits at the PRESENT/ALTERED boundary. We therefore treat the judge as a human-validated but noisy measurement instrument rather than ground truth; Appendix B gives the protocol, the three-way results, and the confusion matrices.

Uncertainty. For longitudinal decay we fit a binomial GEE with passage clusters: survived $\sim C(\text{fact_type}) \times \text{depth_c}$ (Liang and Zeger, 1986). For prompt, operation, and model summaries we prefer fixed-depth survival with passage-level cluster bootstrap intervals, because 82.6% of half-life cells are censored.

4 Results

4.1 Survival decays across rewrite hops

The clean panel contains 339,320 rows. Of these, 3,423 (1.01% of the full clean panel) have unparseable judge outputs and are treated as missing rather than as factual losses, leaving 335,897 measured observations for the GEE. With passage-clustered standard errors over the 25 clusters, each additional hop reduces the odds of survival for the CAUSAL reference category by a factor of OR=0.743 (passage-clustered $p < .001$; pseudo- $R^2 = 0.068$, ICC=0.005). On the probability scale the fitted CAUSAL curve crosses 0.5 at hop 10. Interactions indicate slower decay than CAUSAL for several surface categories and for NEGATION, but not a simple logical-versus-surface ordering. 238/288 half-life cells are right-censored, so fixed-depth survival is our primary endpoint.

4.2 Exact fact-type rankings are unstable

We measure concordance of fact-type rankings across the 4 models using Kendall’s W (Kendall and Babington Smith, 1939). At hop 5, using all 6 fact types, $W = 0.082$ ($p = 0.896$); at hop 12 on the neutral deep subset, $W = 0.027$ ($p = 0.991$). Thus the exact ranking is not stable. In the deep subset, logical facts survive at 0.495 and surface facts at 0.601 at hop 12, but the clean data do not

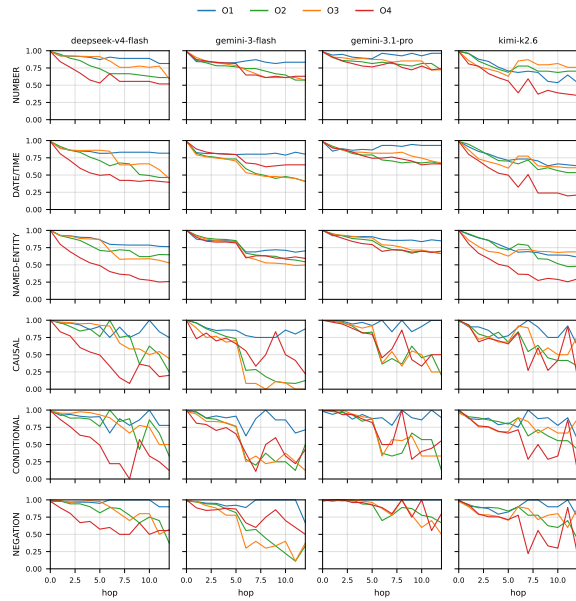


Figure 2: Clean-panel survival curves by fact type and model. Exact type rankings vary across models.

Prompt comparison	Δ survival	95% CI	Passages
Generic faithful - neutral	0.132	[0.084, 0.179]	25
Fact-type-aware - neutral	0.162	[0.118, 0.213]	25
Fact-type-aware - generic	0.026	[-0.022, 0.073]	25

Table 1: Hop-5 prompt effects. Estimates are passage-level cluster bootstrap means; positive values improve over the second condition.

support treating that split as a model-independent total order.

4.3 Prompting helps, but type-awareness adds little

Table 1 reports hop-5 prompt effects. Generic faithfulness improves survival over neutral prompting by 0.132 [0.084, 0.179]. Fact-type-aware prompting improves over neutral by 0.162 [0.118, 0.213]. The fact-type-aware gain over generic faithfulness is small and uncertain: 0.026 [-0.022, 0.073], with an interval crossing zero. In this panel, most of the hop-5 improvement comes from adding an explicit faithfulness instruction rather than from enumerating fact types.

Figure 3 shows the same pattern by fact type: fact-type-aware prompting improves over neutral for all six types, whereas all fact-type-aware-vs.-generic intervals include zero. The estimates therefore support an effect of explicit faithfulness prompting, but not a reliable added effect from naming the fact categories.

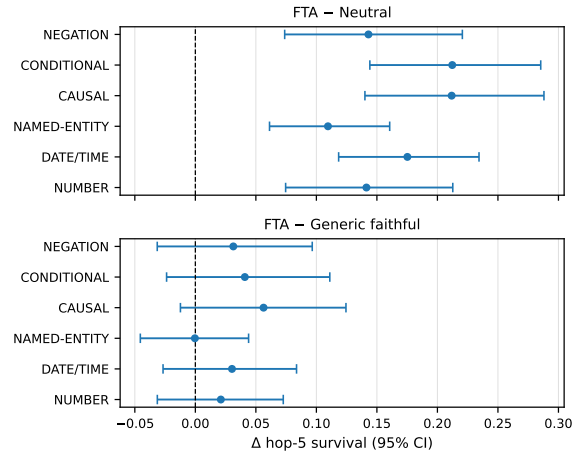


Figure 3: Per-type hop-5 prompt deltas with passage-level cluster bootstrap 95% CIs. Top: fact-type-aware vs. neutral. Bottom: fact-type-aware vs. generic faithfulness. FTA-vs-neutral is positive for all six types; FTA-vs-generic crosses zero for all six.

Deep neutral subset	Hop-12 survival	95% CI
Operation O1	0.752	[0.697, 0.812]
Operation O2	0.571	[0.486, 0.647]
Operation O3	0.586	[0.468, 0.693]
Operation O4	0.466	[0.365, 0.567]
deepseek-v4-flash	0.562	[0.443, 0.638]
gemini-3-flash-preview	0.575	[0.493, 0.654]
gemini-3.1-pro-preview	0.716	[0.664, 0.765]
kimi-k2.6	0.520	[0.422, 0.617]

Table 2: Operation and model survival on the depth-12 neutral subset (8 passages).

4.4 Operations and model identity dominate deep chains

On the hop-12 neutral subset of 8 passages, survival is 0.752 [0.697, 0.812] under direct paraphrase (O1) and 0.466 [0.365, 0.567] under answer-rewrite (O4). The paired passage-level O4-O1 difference is -0.378 [-0.452, -0.300]. The single frontier preview in our panel, Gemini 3.1 Pro Preview, has hop-12 survival 0.716 [0.664, 0.765], while Kimi K2.6 is -0.167 [-0.256, -0.069] lower on the same passages. With one model per tier on this 8-passage subset we cannot separate a tier effect from a provider or version effect; the difference is reported as a between-model contrast on this subset, and the passage-level bootstrap intervals quantify within-subset uncertainty rather than generalization to a broader corpus.

The hop-5 prompt slice compares instruction variants at a fixed depth, while the hop-12 neutral subset measures long-run degradation under the base rewrite prompts. We therefore report generic

faithfulness and direct paraphrase as separate design signals rather than claiming that their gains compose in an unmeasured joint condition.

A single chain illustrates the O4 mechanism. On one Wikipedia passage about 1958 historical events, Gemini 3.1 Pro Preview keeps the named-entity mention *Tristan* alive through the entire O1 chain: hop 0 reads "...after completion of the Lyric Opera of Chicago performances of Tristan und Isolde...", hop 5 "...his final performance, Tristan und Isolde, at the Chicago Lyric Opera", and hop 12 "...leading his final performance, Tristan und Isolde, at the Lyric Opera of Chicago". Under O4 the same model on the same passage drifts into meta-commentary and the mention disappears: hop 5 opens "The provided text is a curated historical summary that highlights notable events, births, and deaths..." and hop 12 "The text you provided is a meta-summary and rephrasing of historical events...". Once O4 has reframed the chain as discussion of the passage rather than a restatement of it, peripheral mentions stop being mentioned at all, and survival collapses even for entities the model recovers at hop 0. The surface matcher is literal about word order: a separate ledger entry, *the Lyric Opera*, scores as lost at hops 1–11 of the O1 chain, where the model writes "the Chicago Lyric Opera".

5 Conclusion

FactRot audits how typed factual mentions persist through iterative LLM rewriting. In this cleaned panel, survival decays with depth, generic faithfulness instructions recover much of the short-chain loss, and deep-chain outcomes depend on rewrite operation and model. The data do not support a stable universal ranking of fact-type fragility, nor a measurable advantage for fact-type-aware prompting over generic faithfulness. Scaling the protocol to larger, snapshot-pinned panels would enable broader benchmark-scale model comparisons.

Limitations

Scale and balance. The cleaned panel contains 1,113 facts from 25 passages, with hop-12 results on 8 passages and one frontier preview model. We therefore interpret model and tier contrasts as within-panel comparisons rather than population-level estimates across model families. The corpus was also deliberately enriched for fact-dense passages covering all six categories, so that the rarer logical types would have enough observations at

this scale; its fact-type frequencies and absolute survival rates should not be read as estimates for Wikipedia text broadly. The hop-5 grid is unbalanced: the frontier preview model was run on the first 15 of the 25 passages, while the three flash-tier models cover all of them. Hop-5 comparisons across models, including the Kendall's W rankings, therefore average the frontier model over a subset of the passages that the other models cover. The hop-12 subset is balanced, with all 4 models run on the same 8 passages.

Measurement noise. Logical-fact measurements remain noisy despite human validation: judge-human agreement is imperfect ($\kappa = 0.516$), with most residual disagreement concentrated at the PRESENT/ALTERED boundary, where human annotators also disagree. Broad trigger lists can select sentences that only superficially express a causal or conditional relation. Surface facts use a strict normalized substring matcher, so aliases, paraphrases, reordered names, and coreferential mentions can count as lost even when the meaning is still recoverable from the text. Half-life is also heavily right-censored (238/288 cells), so fixed-depth survival is the primary endpoint.

Domain scope. All passages come from English Wikipedia. Other domains and other languages may show different fact distributions and rewrite dynamics.

References

- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. [Chain-of-verification reduces hallucination in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhिलाषा Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Bijean Ghafouri and Emilio Ferrara. 2026. [Lost before translation: Social information transmission and survival in AI-AI communication](#). *CoRR*, abs/2602.17674.
- M. G. Kendall and B. Babington Smith. 1939. [The problem of \$m\$ rankings](#). *The Annals of Mathematical Statistics*, 10(3):275–287.

- Kung-Yee Liang and Scott L. Zeger. 1986. [Longitudinal data analysis using generalized linear models](#). *Biometrika*, 73(1):13–22.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Amr Mohamed, Mingmeng Geng, Michalis Vazirgianis, and Guokan Shang. 2025. [LLM as a broken telephone: Iterative generation distorts information](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7493–7509, Vienna, Austria. Association for Computational Linguistics.
- Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. [Generating benchmarks for factuality evaluation of language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian’s, Malta. Association for Computational Linguistics.
- Jérémy Perez, Grgur Kovač, Corentin Léger, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2025. [When LLMs play the telephone game: Cultural attractors as conceptual tools to evaluate LLMs in multi-turn settings](#). In *The Thirteenth International Conference on Learning Representations*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. [Long-form factuality in large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 80756–80827. Curran Associates, Inc.
- Wikimedia Foundation. 2023. [Wikimedia Wikipedia: English wikipedia dump, 2023-11-01](#). <https://huggingface.co/datasets/wikimedia/wikipedia>. Hugging Face dataset configuration: 20231101.en.

A Prompt Templates

The three prompt conditions are stored in the `CONDITIONS` dict in `run_chain.py`. The placeholder `{operation}` is replaced at runtime by the per-operation string from the `OPS` dict (01: paraphrase faithfully; 02: summarize then expand back; 03: write a report from this; 04: answer then restate), and `{text}` is replaced by the output of the previous hop. Line breaks within the templates below are typographic, inserted to fit the column; the literal strings are those in `run_chain.py`.

Neutral.

`{operation}` the following text:

`{text}`

Generic faithful.

Be faithful to the original text. Do not invent, omit, or alter facts.

Now `{operation}` the following text:

`{text}`

Fact-type-aware.

Be especially careful with: NUMBER, DATE/TIME, NAMED-ENTITY, CAUSAL, NEGATION, CONDITIONAL.

Do not invent, omit, or alter any such facts.

Now `{operation}` the following text:

`{text}`

B Judge Protocol and Human Validation

The judge rubric below is the `JUDGE_RUBRIC` string in `survival.py`. The placeholders `{original_text}`, `{current_text}`, and `{bullet_facts}` are filled at runtime with the source passage, the current chain text, and a bullet list of the gold logical facts for that passage.

You are a fact-survival judge.

Original passage:
`{original_text}`

Current chain text (after transformation):
`{current_text}`

Gold facts to check:
`{bullet_facts}`

For each fact above, answer with exactly one of: PRESENT (accurately preserved), ALTERED (changed in meaning or value), or ABSENT (gone).

Format your response as:
VERDICT 1: <PRESENT|ALTERED|ABSENT>

VERDICT 2: <PRESENT|ALTERED|ABSENT>

...

Then give a brief one-sentence overall justification.

Verdict lines are parsed with the case-insensitive regular expression shown below:

```
r'VERDICT\s*<N>\s*:\s*(PRESENT|ALTERED|ABSENT)'
```

A fact survives only when its verdict is PRESENT; ALTERED and ABSENT score 0, and unparseable outputs are treated as missing data. The two automated reliability checks reported in Section 3 are computed from `results/survival_scores_reliability.parquet`.

B.1 Human validation protocol

The validated sample is drawn from the cells the judge scored, in two strata. The 200 uniformly sampled cells form the primary validation stratum. A separate *error* stratum of 60 cells oversamples ALTERED and ABSENT judge verdicts so that the rarer verdicts can be inspected; because it is stratified on the judge's own output, it is used only for error analysis and never pooled with the primary stratum.

Each of the 3 annotators saw the source passage, the gold fact, and the hop text, with the judge verdict, rewrite model, operation, and prompt condition withheld. For the binary endpoint, labels are first collapsed to PRESENT versus ALTERED/ABSENT and then majority-voted, yielding a resolved binary label for all 200 items. The three-way analysis uses the majority verdict and leaves 7 unresolved items out. Confidence intervals resample passages rather than items, since items from one passage are correlated.

B.2 Human validation results

The headline binary agreement is in Section 3. On the 193 main-stratum items with a resolved three-way majority, the judge reaches $\kappa = 0.534$ (agreement 0.876). In the error stratum the human majority confirms 26 of 28 ABSENT verdicts but only 16 of 27 ALTERED verdicts, relabelling 10 of the latter PRESENT. Another 5 error-stratum items had no three-way majority and are omitted from that analysis. The primary-stratum confusion matrix in Table B2 shows the same direction of error; Table B3 reports the resolved error-stratum items. The largest three-way disagreement category is judge ALTERED versus human PRESENT.

Human majority	Judge: not survived	Judge: survived
not survived	20	13
survived	14	153

Table B1. Binary survival, main stratum ($n = 200$).

Human majority	ABSENT	ALTERED	PRESENT
ABSENT	9	0	4
ALTERED	2	7	4
PRESENT	1	13	153

Table B2. Three-way verdicts, resolved main items ($n = 193$); columns are judge verdicts.

Human majority	ABSENT	ALTERED
ABSENT	26	1
ALTERED	1	16
PRESENT	1	10

Table B3. Verdict-stratified error sample ($n = 55$). The judge produced no PRESENT verdicts here by construction.

C Corpus Construction and Fact Extraction

C.1 Corpus construction

The corpus builder (`build_corpus.py`) is deterministic given the snapshot and the seed. It streams the 20231101.en configuration of the `wikimedia/wikipedia` dataset (Wikimedia Foundation, 2023) in an order shuffled with seed 42 and inspects at most 15,000 articles. An article is skipped if its title contains “disambiguation”, if the title begins with “List of”, or if the article is shorter than 300 words.

Within an accepted article, candidate passages are contiguous, sentence-aligned spans: starting from each sentence, sentences are appended until the span reaches 120 words, and the candidate is kept if it is still at most 180 words. Facts are then extracted from the candidate, and candidates yielding fewer than eight facts are discarded.

Candidates are ordered deterministically before selection. Passages containing all six fact types come first, then passages containing the three logical types, then the remainder; within each tier, passages with more facts come first. The first 300 candidates of this ordering are written to `corpus/passages.parquet`. In the study corpus the first tier alone exceeds 300 candidates, so every selected passage contains all six fact types and the file is effectively sorted by descending fact count.

Each extracted span is re-verified against the passage text and dropped if its recorded offsets no

longer reproduce it. Fact identifiers (`fact_uid`) are assigned within this ledger, which is canonical for all reported analyses. Rewrite chains were run on the first 25 passages of the ordering, and the depth-12 subset on the first 8.

C.2 Surface facts

NUMBER, DATE/TIME, and NAMED-ENTITY spans are extracted in `build_corpus.py` by `spaCy` `en_core_web_sm` named-entity recognition, using the fixed label mappings below. Entity labels outside these three sets are ignored.

- NAMED-ENTITY: PERSON, NORP, FAC, ORG, GPE, LOC, PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE.
- NUMBER: CARDINAL, PERCENT, MONEY, QUANTITY.
- DATE/TIME: DATE, TIME.

Two rule-based passes then recover mentions the tagger misses. A date pattern adds month, week-day, ordinal-century and “in YYYY” forms, and a numeric pattern adds integers, decimals, thousands-separated numbers and percentages. The numeric pass skips any span overlapping a NUMBER mention already extracted, so the two sources cannot double-count the same digits. Every retained span is finally re-verified against the passage text and dropped if the recorded offsets no longer reproduce it.

Each mention is stored with a canonical value, and the scorer in `survival.py` applies the same normalization to the hop text before matching, so both sides of the comparison are normalized identically. NUMBER values are lowercased, thousands separators are removed, and `per cent/percent` become `%`. DATE/TIME values are lowercased and stripped of characters outside `[a-z0-9/-]` and spaces. NAMED-ENTITY values are lowercased. All three collapse runs of whitespace. A fact counts as surviving when its canonical value occurs as a substring of the normalized hop text, subject to a digit-boundary guard that prevents a number from matching inside a longer digit string (for example, 540 does not match inside 13540).

C.3 Logical facts

Sentence-level extraction of CAUSAL, NEGATION, and CONDITIONAL facts fires whenever a sentence contains any string from the corresponding list in

build_corpus.py. Counts below include duplicate stems with different trailing punctuation (e.g. "not ", "not,", "not."), which is how the lists are written in the source. Representative samples are quoted here verbatim.

CAUSAL (68 entries). Sample: " caused by ", " causes ", " led to ", " results in ", " due to ", " because ", " since ", " therefore ", " thus ", " hence ", " as a result ", " following ", " after ", " prompted by ", " driven by ". The list is intentionally broad and includes ambiguous temporal markers such as " after " and " following " that can fire on purely chronological sentences without any causal claim. This is the source of the precision limitation noted in the Method section and in Limitations.

NEGATION (88 entries). Sample: " not ", "n't ", "never ", " nor ", "no longer ", "did not ", "cannot ", "lack of ", " failed to ", " refused to ", " rather than ", " without ", " hardly ", " nothing ", " ceased to ".

CONDITIONAL (65 entries). Sample: " if ", " unless ", " provided that ", " as long as ", " given that ", " assuming ", " should ", " were to ", " suppose ", " when ", " whenever ", " in case ", " only if ", " even if ", " depending on ". The list is intentionally broad and includes markers such as " should ", " were ", and " when " that frequently occur in non-conditional contexts (modal advice, plain past tense, temporal clauses); together with the CAUSAL ambiguity above this drives the precision limitation flagged in Limitations.

C.4 Annotated passage examples

The three excerpts below are taken verbatim from the study corpus, with every ledger fact whose span falls inside the excerpt, exactly as the rules produce them.

Example 1 (NEGATION). Passage 4d3464d4, characters 0–95:

John Foster Dulles, the U.S. Secretary of State, responded "We are not afraid of May 27, 1959."

Type	Ledger span
NAMED-ENTITY	John Foster Dulles
NEGATION	whole sentence
NAMED-ENTITY	U.S.
NAMED-ENTITY	State
DATE/TIME	May 27, 1959
DATE/TIME	May 27

The NEGATION fact covers the quoted claim, and the nested surface spans May 27 and May 27, 1959 are kept as separate ledger entries and scored independently.

Example 2 (CAUSAL). Passage c7c214a3, characters 487–684:

William Sands Cox (1802-1875) bought the property in 1873. He was a notable surgeon from Birmingham. He became famous because he founded a prominent medical school and later established a hospital.

Type	Ledger span
NAMED-ENTITY	William Sands Cox
DATE/TIME	1802-1875
DATE/TIME	in 1873
DATE/TIME	1873
NAMED-ENTITY	Birmingham
CAUSAL	whole sentence

The CAUSAL fact spans the sentence carrying the because relation, alongside the person, place, and date mentions of the surrounding text.

Example 3 (CONDITIONAL). Passage 5cc839a2, characters 238–380:

The Elis had a chance to win the Triangular League if it captured both games and both they and the Crimson fought a pitched battle on the 7th.

Type	Ledger span
CONDITIONAL	whole sentence
NAMED-ENTITY	Elis
NAMED-ENTITY	the Triangular League
NAMED-ENTITY	Crimson

The CONDITIONAL fact spans the sentence introduced by if, together with the three named entities it mentions.

D GEE Slopes

Fact type	log-OR/step	OR/step	p_{int}
CAUSAL ref.	-0.298	0.743	1.6×10^{-38}
CONDITIONAL	-0.275	0.760	0.063
DATE/TIME	-0.196	0.822	5.0×10^{-9}
NAMED-ENTITY	-0.217	0.805	2.0×10^{-6}
NEGATION	-0.241	0.786	0.006
NUMBER	-0.182	0.834	8.6×10^{-10}

Table A1. Clean-panel GEE slopes. Non-reference rows combine the CAUSAL depth slope with the fact-type interaction; p_{int} tests the interaction.

E Half-Life Diagnostics

Fact type	Median $h_{1/2}$	Uncens. cells	Censored
CAUSAL	5.5	14/48	70.8%
CONDITIONAL	5.5	12/48	75.0%
DATE/TIME	4.0	9/48	81.2%
NAMED-ENTITY	4.5	6/48	87.5%
NEGATION	6.0	6/48	87.5%
NUMBER	4.0	3/48	93.8%

Table A2. Half-life diagnostics on the clean panel. Heavy censoring makes these secondary to fixed-depth survival.

F Artifacts

The experimental artifacts include the cleaned survival panel, fixed-depth summaries, passage-level bootstrap intervals, judge prompt, trigger lists, human-validation materials, and the numeric audit used to generate the reported results.