From Astronomical Circulars to Auditable Follow-up Records: Schema-Constrained LLM Extraction with Human Verification

Fabian Schüssler[1] **S. Bisero**[1] **M. Cellier**[1,2] **J. Cinosi**[1] **A. Ciric**[1] **B. Cornejo**[1]
I. Jaroschewski[1] **A. Kaan Alkan**[3] **W. Kiendrébéogo**[1] **A. Saint-Paul**[1,2]

[1]Université Paris-Saclay / CEA / IRFU, F-91191 Gif-sur-Yvette, France
[2]École d'ingénieurs aéronautique et spatiale (IPSA), Paris, France
[3]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA
astro.colibri@gmail.com

Abstract

GCN Circulars distribute crucial information on astronomical observations of transient phenomena, but their flexible prose mixes observations, limits, metadata, contacts, and interpretation. We present an auditable pipeline that converts these reports into event-linked, structured records. Regular expressions provide domain hints; a large language model returns strict schema-constrained JSON; deterministic stages normalize values and emit review flags. Seven domain reviewers assessed 231 outputs from 210 distinct Circulars, fully confirming 223 reports and marking eight partly correct. They confirmed 25,827 of 25,880 definite field decisions (99.80%); errors concentrated in timing and facility attribution. Comparison with independent redshift datasets found semantic catalog disagreements but no values misread by our pipeline. Applied to the GCN Circular archive back to 2016, the pipeline has structured 26,811 reports into 68,393 observations, now served in real time by the Astro-COLIBRI platform. We release the extraction pipeline as an open-source package and the derived records as a continuously updated resource.

1 Introduction

Gamma-ray bursts (GRBs) and other transients fade within minutes to days, requiring rapid coordination across worldwide observatories. Many results appear in GCN Circulars, a public rapid-publication channel for time-domain astronomy (Barthelmy, 1998; NASA Goddard Space Flight Center, 2026). A Circular is a natural language report that can mix detections, limits, times, exposures, filters, facilities, redshifts, and contacts in prose or compact tables, making them difficult to aggregate automatically.

Extraction requires more than entity recognition: a number may be an afterglow magnitude, a limit, or calibration-star photometry; a time may be absolute, trigger-relative, or an exposure midpoint. Planned observations and candidate-host redshifts must also be distinguished from completed measurements and spectroscopic redshift determinations.

This paper studies the Circular-to-record component deployed in Astro-COLIBRI (Reichherzer et al., 2021, 2023), whose main flow is shown in Figure 1. Our contributions are:

1. a typed, observation-level extraction task for time-domain reports, each measurement emitted with its time, exposure, band, value, detection status, facility, and a mandatory supporting excerpt (Section 3);
2. a hybrid architecture surrounding a schema-constrained model call with deterministic pre-analysis and post-validation, running in production (Section 4);
3. a field-level audit of the deployed system and an external redshift cross-check against three catalogs, reported with their limits (Section 5); and
4. an open-source release of the pipeline and the derived archive of >68k observations (Section 6).

2 Related Work

Astronomical text. TDAC introduced an annotated time-domain corpus and named-entity-recognition baselines (Alkan et al., 2022); astroECR added coreference, astrophysical relations, and object normalization (Alkan et al., 2024). General-purpose LLM studies found strong prompt sensitivity in multimessenger extraction (Sotnikov and Chaikova, 2023). Closest to this work, Sharma et al. (2026) apply LLMs to the GCN archive for topic classification and redshift recovery. We differ in three ways: the extraction runs on new Circulars

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 17–22
November 10, 2026 ©2026 Association for Computational Linguistics

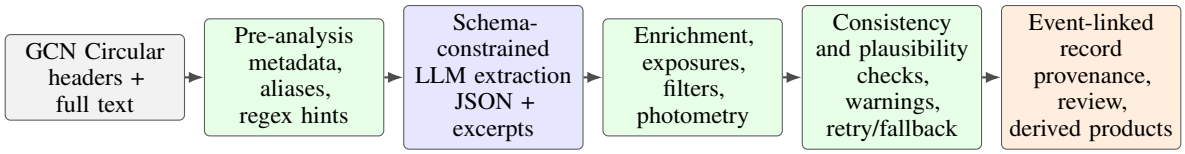


Figure 1: Main architecture. Deterministic stages surround the LLM call, preserve the source and extracted excerpts, and surface ambiguity as warnings. Schema or plausibility failures trigger retry or fallback before an event-linked record is emitted.

in real time and its output is linked to a live event record rather than analyzed offline; the schema is closed and every observation carries a supporting excerpt and full provenance metadata, so each value is traceable to the text that produced it; and we report a field-level human audit of the deployed system. Redshift is where the two efforts touch, and Section 5.3 evaluates ours against three external catalogs.

Scientific information extraction. Document-level scientific IE has long targeted typed records rather than isolated mentions, as in SciERC (Luan et al., 2018) and SciREX (Jain et al., 2020). Dagdelen et al. (2024) showed that LLMs can emit structured JSON records of scientific knowledge directly, which is the template we follow. Our setting differs in that records are consumed within minutes of publication, so an unparseable or implausible output is an ingestion failure rather than a scoring loss.

Constrained generation. PICARD constrains auto-regressive decoding for semantic parsing (Scholak et al., 2021), and grammar-constrained decoding generalizes this to structured tasks without finetuning (Geng et al., 2023). Constraint is not free: Tam et al. (2024) report that format restrictions can degrade reasoning. We accept that trade-off and offset it by keeping the model's job narrow: deterministic code, not the model, performs arithmetic, unit conversion, and cross-field checking (Section 4.3). The mandatory verbatim excerpt makes every field checkable against its source, in the spirit of attribution evaluation for generated text (Rashkin et al., 2023).

3 Task Definition

Given the full text and headers of a Circular, the task is to emit a single closed-schema record. The record contains astronomical source names, a report classification, contacts, an array of observations, and warnings. Report fields cover authorship, facility, instrument, detection, and redshift evidence. Each observation is one measurement or limit and carries time, exposure, band, magnitude or flux, uncertainty, limit status, calibration, notes, and a mandatory source excerpt.

Two properties distinguish this from span annotation. First, the schema is *closed*: additional properties are rejected, so a value that does not fit the record cannot be smuggled into a free-text field, and any output that validates is directly ingestible. Second, the unit of extraction is the observation rather than the mention: a report of six filters yields six rows, each independently timed, calibrated and attributed, and a row is only usable if every one of those fields is correct. Figure 2 shows a Circular fragment and one row it produces.

4 Hybrid Extraction Pipeline

4.1 Advisory pre-analysis

Regular expressions over the subject, headers, and body propose event aliases and locate optical/NIR terms, magnitude-like expressions, redshifts, and contacts. These are advisory prompt hints, never accepted as observations without checking the full text. The stage also normalizes contacts and preserves the unmodified source.

4.2 Schema-constrained model call

One request contains the identifier, subject, hints, and full text; the response must satisfy the schema. The deployment uses `gpt-5.4-mini` with low reasoning effort and `gpt-5.5` as fallback. Transient failures receive two additional attempts per model, while schema or plausibility failures can trigger fallback directly. Model settings are recorded. Fallback was used for 94 of 26,811 analysed reports (0.35%).

Domain rules exclude planned observations and comparison-, calibration-, or host-galaxy photometry. Multi-filter results become separate rows; tables are read by column first, then by row. Repeated exposures retain their notation while count, per-frame duration, and total are normalized. Non-

Input: GCN Circular 40471 (excerpt)

```
The Swift/UVOT began settled observations of
the field of GRB 250516A 108 s after the BAT
trigger (Beardmore et al., GCN Circ. 40467).

Filter    T_start(s) T_stop(s) Exp(s)   Mag
white          108      1715     411  20.5+/-0.3
v              650      1765     136      >18.5
b              576      1691     117      >19.2
u              320      1838     362      >19.7
w1             699      1814     117      >18.8
w2             626      1741      97      >19.1
```

Output: one element of `observations`

```
{"observation_type": "detection",
 "filter_raw": "white",
 "mag_raw": 20.5, "mag_err": 0.3,
 "is_upper_limit": false,
 "observatory": "Swift", "instrument": "UVOT",
 "observatory_short_name": "UVOT",     <-- wrong
 "telescope": "Swift",                 <-- wrong
 "time_raw": "108 s after the BAT trigger",
 "exposure_s": 411,
 "calibration": "Breeveld et al. 2011, ...",
 "source_excerpt": "white  108  1715  411
                    20.5 +/- 0.3"}
```

Figure 2: A Circular fragment containing the complete six-row table and one record it produces. Each table row becomes an independently attributed observation. `source_excerpt` carries the verbatim text supporting the shown row, so every value can be checked against its origin. This example is also one of the eight partly correct reports of Section 5.2: the model identified the facility correctly in `observatory` and `instrument`, but filled the redundant `observatory_short_name` and `telescope` fields with the two values transposed, an error that repeats across all six observations.

optical observations are flagged so their fluxes cannot populate magnitude fields, and every observation quotes its evidence.

4.3 Deterministic enrichment and review flags

Post-processing parses absolute, Modified Julian Date, and trigger-relative times; normalizes filters, facilities, and exposures; and derives observation midpoints from explicit values, intervals, durations, or plausible exposures in that order. Approximations remain flagged. Optical measurements are mapped, when possible, to an observer-frame R_C-equivalent AB display scale while raw values are retained.

Checks flag missing values or times, conflicting representations, implausible dates, pre-trigger observations, and unknown filters. Unresolved candidates and warnings are returned rather than dropped: the LLM handles linguistic variation, while code enforces numerical and interface invariants.

5 Evaluation

5.1 Corpora and protocol

The pipeline was developed on 4,537 Circulars from 2025, with manual inspection of roughly 30 random reports. All 1,775 Circulars published from 2026 January 1 through June 30 completed the held-out operational workflow after configured retries. This measures completion, not accuracy.

Seven reviewers with astronomy and development expertise completed 231 assessments of 210 distinct Circulars from 2020–2026: 115 recent and 95 randomly selected earlier reports, with twenty independently re-reviewed. They compared all extracted fields with the original text, assigned report- and field-level verdicts, and recorded corrections and omissions. Review neither reran the parser nor changed production data.

Field group	Rejected	Reports	Uncertain
Observation time	28	5	7
Facility/instrument	18	4	0
Contact information	4	1	0
Classification	2	1	0
Photometry/limits	1	1	0
Total	53	8	7

Table 1: Rejected and uncertain field verdicts. Content/provenance, filter/calibration, exposure, and redshift had none. "Reports" counts assessments containing a rejection.

5.2 Human-verification results

Reviewers confirmed 223 reports (96.5%), marked eight partly correct, and rejected none. Of 25,887 field verdicts, 25,827 were confirmed, 53 rejected, and seven uncertain: 99.80% of definite decisions were confirmed. One observation was missing.

Timing and facility attribution dominate (Table 1). All 53 rejections occur in the eight partly correct reports, and a single report accounts for 24 of them. Field verdicts are therefore non-independent, making eight of 231 reports the conservative error summary. The twenty re-reviewed reports had unanimous labels and 2,105/2,105 agreeing overlapping field decisions, but contained no rejected fields and thus do not test difficult cases.

That worst case is the example of Figure 2, and it is instructive: its 24 rejections are not 24 independent mistakes but two, each repeated down six rows. The model resolved the facility correctly in `observatory` and `instrument`, then

filled the redundant `observatory_short_name` and `telescope` with the values transposed; and it offset the reported start time instead of taking the midpoint of the reported start–stop interval. A closed schema carrying several near-synonymous fields for one concept invites the first, and our validation checks never compared those fields against each other. Both are cheap to defend against (cross-validate redundant fields, and prefer the interval midpoint when both bounds are present).

5.3 External redshift cross-check

A crucial parameter of astrophysical observations is the redshift (or distance) of the source. Of 468 events with extracted redshifts, 249 appear in GRBweb (Coppin, 2026); 228 (91.6%) agree within $\Delta z = 0.01$, with median absolute difference zero. All 21 residuals were read correctly: nine were limits or conditional associations, seven candidate-host photometric estimates, and five preliminary values. The three reference catalogs themselves agree pairwise for only 96–97% of shared events (Greiner, 2026; NASA/GSFC Swift Science Center, 2026), and our extraction agrees with Greiner for 201/205 events and with *Swift* for 138/139. The comparison supports faithful reading, but shows the schema should distinguish measurements, limits, candidate hosts, conditional associations, and superseded values.

6 Deployment, Scale and Released Artifacts

The pipeline runs on arriving Circulars in real time and has additionally been applied to the archive back to 2016, associating 26,811 reports with 5,787 events, extracting 68,393 observations from 618 facilities, and recovering 84,833 contact records (snapshot: 2026 July 30). Upper limits outnumber detections roughly two to one (42,926 against 20,990), a property of the resource rather than a defect: the non-detections that constrain a fading afterglow are recorded on the same footing as the detections. Of the optical and NIR measurements, 47,515/50,015 (95.0%) reached the normalized plotting scale.

Availability. We release `astro-colibri-circular-parser` v1.0.0 under the BSD 3-Clause license, with a versioned Zenodo archive.[1]

[1] `https://github.com/astro-transients/astro_colibri_circular_parser`; `https://doi.org/10.5281/zenodo.21887735`.

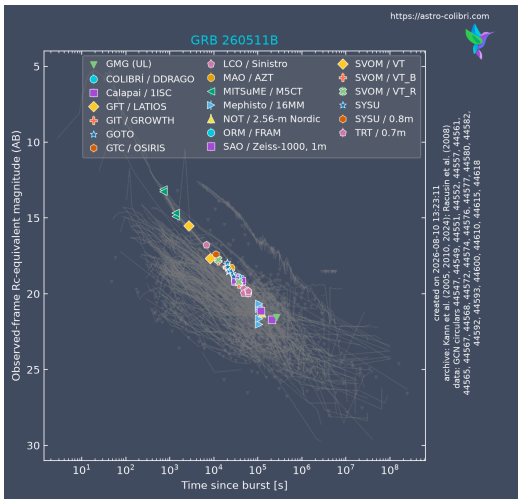


Figure 3: One downstream product: extracted photometry for GRB 260511B (colored markers) placed against archival afterglow measurements (gray). This context figure is intended for rapid inspection, not publication-level photometric analysis.

The package provides Python and command-line interfaces, a cached example for inspecting deterministic stages without an LLM request, and a credential-free test suite. Event association supports the public API, a caller-supplied backend, or offline use without association. Derived summaries and context figures are publicly viewable in Astro-COLIBRI; CSV and IVOA VOTable exports require a user account. The source Circulars are public documents. These products support triage and coordination rather than calibrated catalog analysis; Section 5.2 bounds their reliability.

Downstream, the records drive the optical-afterglow context figure of Figure 3, which plots a live event's extracted photometry against an archival afterglow population on a common observer-frame scale, while retaining links back to the source records and excerpts. The broader deployment and observer-facing products are described by Schüssler et al. (2026).

7 Conclusion

The hybrid pipeline converts free-text Circulars into typed, auditable records by surrounding schema-constrained generation with deterministic guidance and validation. Reviewers confirmed 96.5% of assessed outputs and localized errors mainly to timing and facility attribution. Redshift comparison found minor semantic disagreements but no misread values. Provenance and human verification remain essential.

Limitations

The human verification is an internal precision audit, not an independent benchmark: all reviewers contributed to the project, one supplied approximately 80% of assessments, and sampling emphasized recent reports without stratification by report type or difficulty. Reviewers primarily judged produced fields. Although they could record omissions, recall requires an independently annotated corpus enumerating every target before system output is seen. We therefore report no recall.

Field decisions repeat schema instances and are dominated by larger reports, so they are not 25,880 independent trials. The multiply reviewed subset contained no errors. Redshift validation covers one scalar field, catalogs disagree, and the schema loses whether a value is a limit, candidate-host estimate, or preliminary result.

The proprietary models can be retired or silently updated; recorded identifiers, inputs, and outputs improve traceability but not exact reproduction. Performance is untested for other languages, private reports, other genres, and genuinely multi-source Circulars. The normalized plot uses astrophysical assumptions and is only a context aid.

Ethical Considerations

GCN Circulars are public, but contact details are personal information supplied for scientific communication; deployments should preserve provenance and prevent bulk messaging. LLM output is not authoritative metadata without checks or review. Proprietary inference creates cost, access, and reproducibility concerns, and human users remain responsible for scientific conclusions.

Acknowledgments

The authors acknowledge ANR (French National Research Agency) for its support of the project "Multi-messenger Observations of the Transient Sky (MOTS)" no. ANR-22-CE31-0012. The authors also acknowledge support from the ACME project funded through the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No. 101131928. The authors acknowledge the use of AI-assisted writing and language editing tools during manuscript preparation; all content, interpretations, and conclusions were reviewed and approved by the authors. We thank Pierre Zweigenbaum, Cyril Grouin, and Felix Grèzes for collaboration and discussions on various aspects of the use of AI/NLP for time-domain astrophysics.

References

Atilla Kaan Alkan, Felix Grezes, Cyril Grouin, Fabian Schussler, and Pierre Zweigenbaum. 2024. Enriching a time-domain astrophysics corpus with named entity, coreference and astrophysical relationship annotations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 6177–6188. ELRA and ICCL.

Atilla Kaan Alkan, Cyril Grouin, Fabian Schussler, and Pierre Zweigenbaum. 2022. TDAC, the first corpus in time-domain astrophysics: Analysis and first experiments on named entity recognition. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 131–139. Association for Computational Linguistics.

Scott D. Barthelmy. 1998. The GRB coordinates network. In *Gamma-Ray Bursts, 4th Huntsville Symposium*, volume 428 of *AIP Conference Proceedings*, pages 99–103.

Paul Coppin. 2026. GRBweb: A comprehensive catalog of gamma-ray bursts. Accessed 2026-07-30.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Jochen Greiner. 2026. Table of all well-localized gamma-ray bursts. Accessed 2026-08-07.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

NASA Goddard Space Flight Center. 2026. GCN Circulars. Accessed 2026-06-27.

NASA/GSFC Swift Science Center. 2026. Swift GRB table. Accessed 2026-08-07.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Patrick Reichherzer, Fabian Schüssler, Valentin Lefranc, Julia Becker Tjus, Jayson Mourier, and Atilla Kaan Alkan. 2023. Astro-COLIBRI 2: An advanced platform for real-time multi-messenger discoveries. *Galaxies*, 11(1):22.

Patrick Reichherzer, Fabian Schüssler, Valentin Lefranc, Ali Yusafzai, Atilla Kaan Alkan, Halim Ashkar, and Julia Becker Tjus. 2021. Astro-COLIBRI: The COincidence LIBrary for real-time inquiry for multimessenger astrophysics. *The Astrophysical Journal Supplement Series*, 256(1):5.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901. Association for Computational Linguistics.

Fabian Schüssler, S. Bisero, M. Cellier, A. Ciric, B. Cornejo, I. Jaroschewski, A. Kaan Alkan, W. Kiendrébéogo, and A. Saint-Paul. 2026. AI-assisted extraction of follow-up observations from GCN circulars in Astro-COLIBRI. Preprint, arXiv:2608.23270.

Vidushi Sharma, Ronit Agarwala, Judith L. Racusin, Leo P. Singer, Tyler Barna, Eric Burns, Michael W. Coughlin, Dakota Dutko, Courey Elliott, Rahul Gupta, Ashish Mahabal, and Nikhil Mukund. 2026. Large language model-driven analysis of general coordinates network (GCN) circulars. *The Astrophysical Journal Supplement Series*, 283(1):30.

Vladimir Sotnikov and Anastasiia Chaikova. 2023. Language models for multimessenger astronomy. *Galaxies*, 11(3):63.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.