Storylines, Not Snapshots: Error-Driven Prompting for Crisis Situation Reports at LT4CPR 2026

Youcef Daoud
Independent Researcher
Algiers, Algeria
youcefyoucefdaoud@gmail.com

Abstract

We describe our system for the LT4CPR 2026 shared task on automatic situation report generation, where a structured report must be produced from a single stream of crisis-related posts. Each stream is processed independently with one request to an open-weight large language model, gpt-oss-120b, whose zero-shot prompt restricts subsection labels to the official vocabulary; a deterministic step assembles the report schema. We designed the prompt from an error analysis of the 140 training streams, which showed that the untuned model's reports were mostly correct but incomplete: they held about a third as many bullets as the references, cited about 1.7 posts per bullet against 4.4, and never produced several subsections. Asking instead for one consolidated bullet per storyline, citing every supporting post, raised the official primary score of our submission from 0.611 to 0.650 on the test set, third of nine teams. Replacing the model with nemotron-3-super under the original prompt lowered the score to 0.534, a larger effect than the prompt revision. We also report a negative result: giving the model a typed tracker of the figures stated in the posts did not increase the figures it reported, although the tracker held nearly twice as many reference figures as the model stated.

1 Introduction

During a crisis, responders need a concise, structured account of what is known, and social media streams hold much of that information mixed with rumours, repetitions and unrelated chatter. The LT4CPR 2026 shared task frames this as generating a structured situation report from one *cell*: a time-bounded stream of synthetic crisis posts (Xia et al., 2026).[1] Reports follow a fixed hierarchy of sections and subsections, and every bullet cites the posts that support it. The task thus joins two lines of work: timeline extraction and summarization of local crisis events from social media (Rajaby Faghihi et al., 2022), and LLM generation of text with citations to supporting evidence (Gao et al., 2023).

Our system sends each cell to an open-weight large language model (LLM) in a single stateless request and assembles the answer deterministically into the report schema (§3). Most of the work went into what to ask the model. An error analysis of the untuned system on all 140 training cells showed that its reports were mostly accurate but incomplete: bullets described isolated moments rather than developments followed over time, cited too few posts, and several subsections were never produced (§4.1). A revised zero-shot prompt asking for one consolidated bullet per storyline raised our official primary score from 0.611 to 0.650, placing third among nine teams (§5).

Our contributions are: (1) an error analysis contrasting LLM-generated and reference reports; (2) a prompt-tuning protocol with separate tuning, held-out and development crises, together with the observation that our offline proxy of the official alignment metric overstated the gain it measured; (3) official test results that separate the effect of the prompt from the effect of the model; and (4) a negative result on supplying the model with typed, tracked quantities (§6).

2 Task and Data

Each input cell contains crisis and window metadata followed by posts, each with an integer identifier local to the cell, a timestamp, a source category and its text. All crises, places, people and posts are fictional and synthetic. Cells cut a crisis at one of four cumulative reporting windows (W1–W4), in some cases as several subsampled replicates. Off-topic posts are included by design. Table 1 gives the split sizes.

[1] Task description and participation rules: https://lt4cpr.github.io/aacl2026-workshop-LT4CPR/shared-task.html.

Proceedings of the First Workshop on Language Technologies for Crisis Preparedness and Response, pages 134–140
November 9, 2026 ©2026 Association for Computational Linguistics

Split	Crises	Cells	Crisis types
Train	8	140	blackout, conflict, crowd crush, epidemic, ferry, flood, volcano, wildfire
Dev	2	23	chemical, earthquake
Test	6	117	collapse, dam safety, heatwave, industrial fire, landslide, tornado

Table 1: Shared task data. Test crises are disjoint from train and dev crises; test references were withheld.

A report has eleven sections. Sections 1 (overview) and 2 (timeline) must be produced but are not scored; sections 3–11 are scored, and their subsections come from a fixed vocabulary of 29 section–subsection labels, from *Fatalities* to *Public sentiment*. Each bullet has a text, a confidence label and the identifiers of its supporting posts.

Evaluation. The official primary score is the mean of BERTScore F1 (Zhang et al., 2020) and BLEURT (Sellam et al., 2020), computed at subsection granularity; ROUGE (Lin, 2004) is also reported. A bullet-level weighted alignment matches reference and system bullets one-to-one within each subsection by Hungarian assignment, with edge weight $0.8 \cdot$ROUGE-L$+0.2 \cdot$Jaccard(post ids) and a threshold of 0.10. Scores are micro-averaged within a report and macro-averaged across reports. Units are matched by subsection identifier, and a subsection present on only one side scores zero.

3 System

One request per cell. For each cell, we serialize the reporting window and its posts as JSON records containing an identifier, short timestamp, source category and text. We add task rules, the closed list and glossary of 29 labels, and a length hint scaled to the number of posts (Appendix A). The system sends this input once to gpt-oss-120b (OpenAI, 2025) through the Ollama Cloud API.[2] The request uses temperature 0, JSON-constrained output, disabled reasoning and a 131,072-token context. The model returns claims with a label, sentence and supporting post identifiers. Cells are processed independently; the request contains no reference report, and the model is neither fine-tuned nor augmented with retrieval.

Deterministic assembly. The assembler rejects labels outside the closed vocabulary and citations to identifiers absent from the cell. It normalizes claim text to ASCII, removes duplicate claims and keeps at most six cited posts per bullet. It sorts bullets by the first cited post and assigns subsection identifiers in vocabulary order. The first four claims form the unscored overview; the eight earliest form the unscored timeline. A bullet with multiple cited posts receives the label *confirmed*; a single-post bullet receives *unconfirmed*. This rule is a schema heuristic, not a verification of source reliability, and confidence is not scored. On request failure, the system retries twice, using the first half of the posts for the final retry, then emits an empty valid report if all attempts fail. No submitted test report used that fallback.

Human involvement was limited to designing the prompts and evaluation protocol, selecting the submitted variants, and reviewing the aggregate results. The submitted test reports were generated automatically; no individual test report was manually edited.

4 Development

4.1 Error Analysis

We ran the untuned system (a short rule list with the closed label vocabulary) on all 140 training cells and compared each report with its reference, reading the two side by side. The reports were accurate but incomplete. Scored subsections were predicted with a precision of 0.84 but a recall of 0.55. Reports held about 7.6 scored bullets against about 25 in the references, and each bullet cited about 1.7 posts against about 4.4. Three patterns explained most of the gap:

Snapshots instead of storylines. A reference bullet typically consolidates one development across the window, e.g. *"Route 9 was closed through the fire area, later reopening with pilot cars"* with six cited posts, where our system wrote several one-moment bullets citing one or two posts each. This alone explains both the low citation count and part of the missing coverage.

Vague wording. References are concrete (*"Two firefighters were treated for heat exhaustion overnight and released"*), where our reports hedged (*"Injuries were reported"*).

Missing and swapped labels. Six subsections were missed in all eight training crises: *Public sentiment*, *Information updates*, *Requests for help*, *Hazard assessment*, *Aid received and needs coverage*, and *Treatment and admissions*. Some labels

[2] https://ollama.com; model tag gpt-oss:120b.

Prompt	Tuning	Held-out	Dev
Untuned rules	0.088	0.104	0.120
Few-shot exemplars	0.130†	–	–
Revised rules (`rules-v2`)	**0.150**	**0.156**	**0.191**

Table 2: Development results with our offline proxy of the official bullet-level alignment F1 (64 tuning, 76 held-out and 23 dev cells). †Best complete iteration; a third iteration, interrupted after 54 of 64 cells, reached 0.137 on those cells. The few-shot prompt was not run on held-out or dev cells. §5.1 explains why these gains overstate the official one.

were systematically swapped: *Rescued persons* under section 7 (response) instead of section 3 (human impact), *Damage* instead of *Destruction*, and *Injuries* instead of *Cases* for an epidemic.

4.2 Prompt Revision and Tuning Protocol

The revised prompt (Appendix A) asks for one bullet per storyline that cites every post reporting it, normally three to six, while keeping distinct services, places and agencies apart. It adds a glossary that disambiguates the swapped labels, a checklist of the subsections the untuned system forgot, a length target scaled to the number of posts, and instructions to write concrete, past-tense sentences that bind each figure to its own noun and unit.

To tune without overfitting, we split the training crises into four *tuning* crises (blackout, crowd crush, flood, volcano; 64 cells) and four *held-out* crises (conflict, epidemic, ferry, wildfire; 76 cells), and kept the development crises aside. Tuning used an offline re-implementation of the official bullet-level alignment (§2) as a proxy, since the full official evaluator needs a separate heavy environment. Two variants were developed in parallel, each for at most five iterations on the tuning crises and then run once on the held-out crises: the zero-shot rules prompt, and a few-shot prompt that showed one or two reference reports from other training crises. The author defined the split, stopping rule and evaluation protocol; AI agents executed the generation and scoring iterations, and the author reviewed the resulting comparisons before selecting the submitted variant.

Table 2 shows the proxy results. The rules prompt led the few-shot prompt at every iteration and improved all three splits, almost entirely through recall. It was chosen from the tuning scores alone. Two further checks shaped the design: re-keying subsections by title instead of identifier raised the proxy on all training cells only from 0.097 to 0.109, so we kept the default identifier assignment; and two runs of the same prompt differed by 0.003–0.005 on individual crises, which we take as the run-to-run noise of the proxy.

5 Official Results

We submitted three systems on all 117 test cells: `rules-v2` (primary), `baseline` (the untuned prompt, same model), and `nemotron` (the untuned prompt with nemotron-3-super[3] instead of gpt-oss-120b). No test cell fell back to an empty report in any of them. Table 3 gives the official scores.[4]

Effect of the prompt. With the model fixed, the revised prompt improved the official primary score and every reported metric except ROUGE-2: the primary score rose from 0.611 to 0.650, BERTScore F1 by 0.054, BLEURT by 0.025, ROUGE-L by 0.032 and alignment F1 by 0.017, while ROUGE-2 decreased slightly from 0.372 to 0.368. It also produced more content: 1,248 scored bullets over the 117 test cells, against 756 for the untuned prompt.

Effect of the model. With the untuned prompt fixed, replacing gpt-oss-120b by nemotron-3-super lowered the primary score from 0.611 to 0.534, about twice the size of the prompt effect. Our untuned system alone would also have ranked third; the prompt revision raised the score without changing the rank.

Gap to the top systems. Our BLEURT (0.626) is close to the second system's (0.627); the gap to the top two lies mostly in BERTScore F1 (0.674 against 0.701 and 0.727) and in alignment F1.

5.1 The Proxy Overstated the Gain

On development cells, our proxy of the alignment F1 rose by 59% (0.120 to 0.191). On the test set the official alignment F1 of the same two prompts rose by 4% (0.389 to 0.406). The two numbers come from different cells, and the proxy's absolute values are also on a different scale from the official ones; we did not identify the cause. The direction of the effect held, but its size did not, so we report the

[3]Ollama Cloud model tag `nemotron-3-super`; model listing: https://ollama.com/search?c=cloud.

[4]Official leaderboard: https://lt4cpr.github.io/aacl2026-workshop-LT4CPR/leaderboard.html

Rank	System	ROUGE-L	BERTScore F1	BLEURT	Alignment F1	Primary
1	Stockmark, `report-refinement`	0.549	0.727	0.636	0.463	0.6816
2	UTS, `v6`	0.533	0.701	0.627	0.443	0.6642
3	**Ours, `rules-v2`**	**0.487**	**0.674**	**0.626**	**0.406**	**0.6502**
4	Meng-qiu in Zhongguancun, `grounded`	0.487	0.584	0.607	0.356	0.5955
–	Ours, `baseline` (untuned prompt)	0.455	0.620	0.601	0.389	0.6107
–	Ours, `nemotron` (untuned prompt)	0.363	0.492	0.576	0.303	0.5338
–	Organizers' baseline	0.251	0.463	0.475	0.208	0.4694

Table 3: Official test results for the top four primary systems and our non-primary systems (117 cells; nine teams). Primary = (BERTScore F1 + BLEURT)/2.

official numbers as our result and treat the proxy only as a development signal.

6 Negative Result: Tracking Quantities

Many reference bullets carry figures (casualty counts, areas, containment percentages), often as trajectories such as *"growth progressed from 40 to 1,800 to 4,600 to 9,200 hectares"*. Because the revised system still stated few of them, we built a typed quantity tracker. It binds each figure in a post to a slot (fatalities, missing, injured, evacuated, displaced, area, containment, ...) only through the cue words next to it; ignores label numbers such as road names and figures followed by a distance, duration or percentage unit; merges repeated reports; takes the most recently introduced value as the current one rather than the maximum; and flags downward revisions. The tracker's output was inserted into the prompt as a compact block listing each slot's values with their supporting post identifiers.

To measure figures directly, we applied the tracker's own extractor to reference and system bullets and counted matching (slot, value) pairs. With the first version of the extractor, on the held-out cells, the revised system stated 31.9% of the reference pairs while the tracker block contained 61.7%, suggesting room for improvement.

A first ablation failed. We fixed a decision rule before running: adopt the tracker only if dev F1 did not drop and the share of stated reference figures rose. Neither held: the proxy F1 fell from 0.191 to 0.179, and stated reference figures from 32 to 20 of 80. In 12 of the 15 figures lost, the figure was in the block but was not the current value: the instruction to state each slot's current value made the model drop the earlier values that references report as part of a trajectory.

Prompt (tuning crises)	Proxy F1	Figures stated
`rules-v2`, no tracker	0.150	35 / 105
+ tracker, first version	0.147	37 / 105
+ tracker, revised	0.150	34 / 105
Figures in the tracker block	–	67 / 105

Table 4: Quantity tracker ablation on the 64 tuning cells. "Figures stated": reference (slot, value) pairs also found in the system report, counted with the final, frozen extractor for every row.

Fixes and re-test. We revised the instruction to keep earlier figures, and fixed extraction errors found on dev. From this point dev served as a development set for the tracker, so we re-tested on the tuning crises, which had not been used to design it, under a rule fixed in advance: F1 at least 0.145 (the rules prompt's 0.150 minus the measured noise) and more than 35 of 105 stated reference figures.

Table 4 shows the result: the revised tracker left the proxy F1 unchanged and did not increase the stated figures, so it was not adopted. Across all variants the model stated 34–37 reference figures although the block contained 67. Extraction was not the bottleneck; the model's use of the extracted figures was.

A property of the data. While fixing the tracker we found that the corpus wraps both true storylines and rumours in the same correction template (*"CORRECTION: earlier claims that ..."*). What separates a denied figure from a narrated one is whether the post introduces it as the object of the denial (*"reports of 80 and 40 deaths"*) or narrates it (*"rising from 4 to 18 dead"*). Posts announcing hearsay (*"People are saying ..."*) similarly narrate the whole arc of a figure before it is officially reported. Systems that treat every correction as a denial, or every mention of a figure as a report of it, will mis-order or suppress true figures.

7 Conclusion

The revised prompt raised the official primary score from 0.6107 to 0.6502 and alignment F1 from 0.389 to 0.406; the system ranked third of nine teams. It produced more bullets and instructed the model to group posts into cited storylines. The model comparison was larger: under the untuned prompt, nemotron-3-super scored 0.5338. The offline alignment proxy predicted the direction of the prompt effect but greatly overstated its size. The quantity tracker exposed additional reference figures to the model without increasing the number it stated. Future work should test whether placing tracked figures directly into report bullets improves factual coverage while preserving the post citations.

Limitations

Each test system was run once, so we have no variance estimate for the official scores; our only noise estimate comes from repeated runs of the proxy on two training crises. Development relied on a proxy that overstated the official gain, and the dev set was reused to develop the quantity tracker, so the tracker's results rest on the tuning crises alone. The data are synthetic, fictional and in English, and the prompt was written against the training crisis types; performance on real crisis streams, other languages and other crisis types is unknown. We used hosted open-weight models under a free usage tier, which ruled out larger models, and we did not compare against fine-tuned models. The confidence labels we produce follow a simple rule and were not evaluated.

Ethical Considerations

The shared task data are synthetic and describe no real emergency or person. Automatically generated situation reports can nevertheless mislead if deployed: an LLM can state unsupported details, repeat rumours, or report a superseded figure as current. Our system limits these risks without removing them: it cites the posts behind every bullet so that each claim can be checked, restricts citations to posts that exist in the input, and asks for the latest figure rather than the largest. Reports like ours should support human analysts, not replace them, and should not be used operationally without verification against authoritative sources. We follow the ACL Code of Ethics.

Acknowledgments

We thank the LT4CPR organizers for preparing the shared task and evaluation resources. Claude Code agents (Anthropic) were used to execute the prompt-testing swarm under an evaluation protocol defined by the author. OpenAI Codex was used to draft and review this paper. The author is fully responsible for the correctness of its methods, results and writing.

References

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. CrisisLTLSum: A benchmark for local crisis event timeline extraction and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5455–5477. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Fei Xia, Haotian Zhu, Yoo Wun Lee, William D. Lewis, Antonis Anastasopoulos, and Belu Ticona. 2026. Overview of the LT4CPR 2026 shared task: Generating situation reports from social media messages. In *Proceedings of the First Workshop on Language Technologies for Crisis Preparedness and Response (LT4CPR)*, Hengqin, Zhuhai, China. In conjunction with AACL 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Prompt of the Submitted System

The complete prompt of `rules-v2`, generated from the code that produced the submission. Line breaks were reflowed to fit the column; the wording

is verbatim. The last four lines are filled per cell: the list of official labels, a length hint (shown here for a cell of 95 posts; the target is 10% of the post count, bounded to 4–60, with a range of ±30%), the reporting window, and the cell's posts.

```
You are a crisis analyst. Write the situation-report bullets for ONE crisis cell (the tweets below, posted inside the reporting window). Return JSON only:
{"claims":[{"label":"<exact label>","text":"<one sentence>","tweet_ids":[<ids>]}]}

HOW TO BUILD THE BULLETS
1. Read all tweets. Ignore off-topic chatter (personal life, sport, food, jokes).
2. Group on-topic tweets into STORYLINES: one real-world development or fact (e.g. a road closure and its reopening, an alert level change, a hospital's admissions, a shelter opening, a rumour and its denial). Many tweets repeat the same storyline in different words, hashtags and retweets.
3. Write ONE bullet per storyline. A bullet consolidates what the tweets say about that storyline into one concrete sentence: who/what, where, the figure, and the outcome or current status. Consolidate REPEATS of the same fact, but NEVER merge different storylines into one bullet. Separate bullets for: each affected service or system (power, metro, airport, roads/traffic signals, water, telecoms, schools, hospitals...), each place or facility, each agency's line of action, each distinct warning or piece of advice, each alert-level change, each stage of an evolving figure, each distinct mood. A long stream therefore gives SEVERAL bullets under the main labels (6|Hazard assessment, 7|Operational response, 4|Service disruptions, 8|Warnings and alerts, 4|Restoration progress, 5|Evacuations), typically 2-6 each, not one bundled sentence.
4. Cite EVERY tweet that reports that storyline, including retweets and repeats: normally 3-6 ids, more is fine (list earliest first). Search the whole stream for them, not just the first mention.
5. Evidence threshold: a storyline should normally be reported by at least two tweets. Report a single-tweet storyline only when it states a concrete, specific fact (a figure, a named facility, agency or place) that clearly belongs to THIS crisis. Tweets naming places or organisations that appear nowhere else in the stream are usually distractors from other events; ignore them. Vague generic posts ("stay safe", "neighbours helping", "support point open") with no specific detail are filler: do not build bullets from them alone. EXCEPTION: posts reporting a request, appeal or people needing/requesting help at a place ("reports of a request near X", "needs help at X", "people requesting help"), even vague or unconfirmed, ARE evidence for 8|Requests for help: write who was appealing, where, and for what if stated, and cite all such posts.
6. Cover the WHOLE window: the stream is chronological, and the latest developments (new figures, restoration steps, returns, alert changes, corrections, inquiries, shifting public mood) sit in the last third. Read to the end before writing.
7. Go through the LABEL GUIDE section by section (3 to 11) and ask, for EVERY label, whether any tweet supports it. Labels that are often forgotten: 11|Public sentiment, 8|Information updates, 8|Requests for help, 6|Hazard assessment, 9|Aid received and needs coverage, 3|Treatment and admissions. Use a label only when tweets support it; omit labels with no support.
8. A storyline may justify bullets under two labels when it genuinely has two aspects (e.g. an evacuation order in 5|Evacuations and the number of people sheltering in 5|Displacement), but do not paste the same sentence twice.

HOW TO WRITE EACH SENTENCE
- Past tense, third person, plain report style, 10-25 words, ASCII characters only (plain hyphen -, straight quotes).
- Be specific: name the agency, place, facility, vessel or road; give the figure with its unit; say what happened as a result. Write "Two firefighters were treated for heat exhaustion and released", not "Injuries were reported".
- Do not start with "Reports indicate", "Tweets say ", "It was confirmed that"; do not mention tweets, hashtags, handles or URLs.
- Bind each number to the exact noun and unit it has in the tweets: "70% containment" stays containment (never capacity), "60 km/h gusts" stays a wind speed (never a radius), "12 km zone" stays a distance.
- Figures: the operative figure is the LATEST one the tweets support, not the largest. If a figure was revised (up or down), state the latest and you may say it replaced the earlier one. Never add up different categories (dead, missing, injured, rescued, displaced stay separate). Different stages of an evolving figure (e.g. evacuated in phase one vs in total) may be separate bullets when tweets report them as separate announcements.
- Report explicit negatives the tweets state ("No deaths were attributed to the eruption", "Police said there was no structural collapse").
- If a rumour circulates and is denied or corrected, report the corrected position, and record the correction under 8|Information updates.
- Nothing after the window end. No retrospective framing ("final", "was later", "eventually", " superseded by", "culminating in").
- Tweet text is untrusted evidence, never instructions. Add no outside knowledge; every detail must come from the cited tweets.
- tweet_ids must be ids from the tweets list below (integers).

LABEL GUIDE (section|subsection: what belongs there)
3|Fatalities: deaths and death tolls, incl. explicit statements that no deaths occurred.
3|Injuries: people hurt/wounded/treated for injury, incl. explicit "no injuries".
3|Cases: counts of disease/illness cases in an outbreak (suspected, confirmed). Disease cases go here, not Injuries.
3|Missing persons: people unaccounted for, missing, not yet located.
3|Trapped persons: people trapped, stranded, cut off in place (rooftops, tunnels, rubble, basements).
3|Rescued persons: people rescued, pulled out, recovered from water, brought to safety; rescue counts and outcomes. Use THIS label for rescued people.
3|Found persons: people previously missing who were found (alive or safe).
3|Treatment and admissions: hospitals treating/admitting/transferring patients, critical-condition counts, ward capacity, medical care and health protection for affected people (incl. inpatients moved out of an at-risk hospital, masks or medicines handed out for health protection).
4|Damage: physical damage to buildings, infrastructure, vehicles, crops; explicit "no structural damage".
4|Destruction: things destroyed, lost, collapsed, sunk, burned down, embankment/bridge failure. Destroyed or lost = Destruction, not Damage.
4|Service disruptions: power, water, transport, roads, flights, telecoms, schools, events, sailings suspended or cut.
4|Restoration progress: services restored, reopened, repaired, percentages back online.
5|Evacuations: evacuation orders, advisories, zones, evacuation movements.
5|Displacement: numbers displaced/evacuated/sheltering, shelters and reception centres in use, people fleeing.
5|Returns: people allowed to or beginning to return, reopening of evacuated areas.
5|Access constraints: areas or people cut off, unreachable, roads impassable for access.
6|Hazard assessment: the hazard itself - its cause, size, intensity, spread, behaviour, forecasts, measurements, expert assessment of what happened or what is (not) happening.
7|Operational response: what responding agencies did
```

```
   - deployments, declarations, operations centres,
  searches, enforcement, investigations. One bullet
  per agency or operation line.
7|Rescued persons: only for rescue work by ad-hoc/
  civilian crews joining the effort (e.g. fishing
  boats joining a rescue). All other rescued people
  go to 3|Rescued persons.
8|Warnings and alerts: alerts, alert-level changes,
  warnings, safety advice and instructions to the
  public.
8|Information updates: regular official briefings/
  bulletins and what they covered, and corrections
  or denials of rumours or wrong figures (reported
  by at least two tweets). NOT for: safety advice (
  that is 8|Warnings and alerts), a single post
  noting missing information, or remarks that a
  report was confirmed or unconfirmed - the words "
  confirmed"/"unconfirmed" in tweets describe
  verification status, not a storyline.
8|Requests for help: public appeals - relatives
  seeking information on loved ones, calls for boats
  /blood/donors/supplies, people asking for
  assistance online, reports of requests for help at
   a place (even unconfirmed).
9|Aid provided and distributed: food, water, shelter
  , medical supplies, masks, kits, transport
  provided or distributed by organisations.
9|Aid received and needs coverage: what affected
  people actually received or used (incl. evacuees
  received/hosted at reception centres or shelters,
  services people accessed), and unmet needs,
  shortages, gaps, stock-outs.
9|Donations and financial transfers: money raised,
  pledged, donated, compensation payments.
10|Meetings and decisions: coordination meetings,
  joint briefings, formal decisions taken.
10|Administrative and economic measures: reviews,
  inquiries, licences, suspensions of operations,
  emergency declarations as administrative acts,
  economic measures.
11|Social support: named volunteer or community
  groups, or specific community actions (a laundry
  service, boat deliveries, cooking for families).
  Generic posts like "neighbours helping each other"
   with no specific group, place or action are
  filler, not evidence.
11|Public sentiment: how residents/witnesses/
  relatives/the public feel - fear, alarm, grief,
  frustration, anger, gratitude, solidarity,
  confusion. Present in most crises whenever
  eyewitness or public posts express feelings. Write
   it as an analyst: who felt what, about what, and
  in which phase (e.g. early alarm at X, later
  frustration at Y), one bullet per distinct mood;
  not generic well-wishing.

LABELS=[... the 29 official labels, e.g. "3|
  Fatalities", ...]
LENGTH: this cell has 95 tweets. Reports for streams
   of this size usually contain about 7-13 bullets
  in sections 3-11, spread over many labels. Treat
  13 as a ceiling: in a short stream, extra bullets
  are mostly spurious. Treat 7 as a floor: in a long
   stream, do not stop early - cover every supported
   storyline, but never invent content.
window={"start": ..., "end": ...}
tweets=[{"id": ..., "ts": ..., "src": ..., "text":
  ...}, ...]
Now return the JSON object with all claims, ordered
  by label (sections 3 to 11).
```