

Q4 - Real first-week extraction experiment

Algebris Data Analytics Case Study | Shi Yuchen | 18 October 2026

Results

A real DeepSeek run processed all **14 morning/evening CTP-ISW Iran Updates from 28 February through 6 March 2026**. All **520 narrative paragraphs** were covered by **71 accepted chunks**, with zero missing paragraphs. The final model outputs contain **76 event mentions**; automated semantic review retains **59**, excludes **17**, and annotates/corrects **58** of the retained mentions. These are report mentions, not independent incidents.

The workflow archives articles, extracts structured events and records each validation and review decision. Exact quotes can still be misinterpreted, so the reviewed analysis and original model responses are stored separately. The semantic review used AI; human verification and a held-out precision/recall evaluation remain outstanding.

Corpus, method and coverage

The corpus includes both daily editions throughout the seven-day window. The [official Iran Updates directory](#) supplied 13 links; the omitted [4 March evening report](#) was found on its official page and included. The reports state that CTP and ISW jointly publish the series. The archive contains rendered DOM exports, source URLs, visible publication dates, actual retrieval timestamps and SHA-256 hashes.

Articles were collected on **18 October 2026**. This comparison uses publication dates to describe the first week; later page revisions may be present, so it cannot reconstruct a contemporaneous trading information set.

The parser selects article narrative, keeps every paragraph in order, and separates the bibliography without duplicating it into the model input. The source audit verifies 309,100 narrative characters and 1,393 reference IDs (including one alias). Original reference links remain attached to their paragraphs. Coverage is limited to narrative text. Images/maps were not OCRed, and linked sources were not fetched or independently verified.

Publication date	Articles	Paragraphs	Chunks	Raw mentions	After review
2026-02-28	2	66	8	4	2
2026-03-01	2	68	11	11	7
2026-03-02	2	82	12	22	18
2026-03-03	2	94	11	14	10
2026-03-04	2	83	12	14	12
2026-03-05	2	78	9	9	8
2026-03-06	2	49	8	2	2

deepseek - f l a s h was called with thinking disabled, JSON-object mode and the full Pydantic schema in the system message. Local Pydantic and exact-evidence checks run independently of provider formatting. The API returned the model alias deepseek - f l a s h; its underlying weights were not independently pinned. See [DeepSeek API documentation](#).

Each paragraph belongs to exactly one chunk; the 12,000-character target includes serialized paragraph IDs and source links. Oversized paragraphs fail explicitly. The batch preserves raw responses, per-attempt usage, request IDs, versions and failure traces. Accepted chunks can be replayed offline without further

model calls. Cross-chunk context loss remains possible; summaries and body paragraphs can repeat an incident.

Validation and review

Pilot runs included generic military reports and misattributed some claims. The final prompt, `conflict-inflation-3`, requires a direct economic connection and separates threats, attacks, damage and operational interruption. Pilot outputs remain in the archive but are excluded from the final analysis. Because these reports informed prompt changes, the corpus is not a held-out benchmark.

Two chunks initially failed field or exact-quote checks. Clearer validation feedback and a rerun recovered both while reusing accepted chunks. The original failures remain in `quarantine_initial/`; quote-matching standards stayed unchanged.

Automated semantic review compared all 76 final mentions with their cited paragraph context. Examples of issues included a vessel's position being confused with an explosion's distance from that vessel, an attack being labelled physical energy-asset damage without supporting damage evidence, a reporting date being treated as an event date, and a military-port incident without a supported commercial effect. The review also flags or excludes clearly redundant summary/body mentions. `q4_semantic_audit.json` records each decision and explicit overrides; `raw_events.csv` and original responses are preserved separately from `reviewed_events.csv`.

Complete paragraph coverage does not measure extraction recall. Production evaluation needs human labels on unseen reports, checks of attribution and quantities, and event resolution across repeated mentions.

How the reported channels changed

2026-02-28 - Warnings and voluntary avoidance precede evidence of a military blockade. The evening report attributes a 70% decline in Hormuz commercial vessel traffic to MarineTraffic and reports an IRGC warning via an EU official. This is a traffic measure, not a 70% loss of global oil supply; the report did not establish a blockade. Sources: [2026-02-28 evening](#).

2026-03-01 - Reports move from warnings to attacks on vessels and commercial ports. UKMTO reports incidents, while Iranian officials deny an intention to close Hormuz at that time. The reported Persian Gulf and Hormuz traffic declines refer to different locations/dates; they cannot be spliced into one measured series. Insurance and transit-cost claims remain attributed. Sources: [2026-03-01 morning](#), [2026-03-01 evening](#).

2026-03-02 - Reported production interruptions broaden the shipping-risk channel: QatarEnergy LNG suspension, the precautionary Ras Tanura refinery shutdown and Kurdish oil-field stoppages. Ras Tanura's precautionary closure does not establish physical damage. Qatar's LNG share describes exposure; it does not quantify the supply loss. Sources: [2026-03-02 evening](#).

2026-03-03 - Downstream and storage constraints emerge: QatarEnergy announces stoppages including urea, while the report connects Rumaila's closure to storage filling as tankers struggle to leave. These reports add fertilizer and export/storage channels. US political-risk insurance and possible tanker escorts are proposed mitigations; any reduction in freight or insurance costs remains to be measured. Sources: [2026-03-03 morning](#), [2026-03-03 evening](#).

2026-03-04 - Maersk booking suspensions widen logistics restrictions, with specified port exceptions. The report cites an approximately 90% Hormuz vessel-traffic decline and a forecast of at least one month for Qatar to regain normal LNG output. The latter concerns recovery to normal output, not a month of zero production; traffic counts do not measure lost energy volumes. Sources: [2026-03-04 evening](#).

2026-03-05 - Reported fires at Sarsang and BAPCO add concrete facility incidents, while the Sonangol Namibe case extends merchant-vessel risk into the northern Gulf. Saudi rerouting through Red Sea

pipelines and continuing Iranian cargo/oil movements indicate partial substitution and residual flows. Regional export flows therefore continued in part. Sources: [2026-03-05 evening](#).

2026-03-06 - UKMTO reports an attack on a tug assisting Safeen Prestige; the evening report also describes attacks on oil-related targets around Basra without confirmed physical damage. Maritime and energy risks persist, but fewer extracted mentions on this day cannot establish de-escalation, recovery, or a lower future CPI print. Sources: [2026-03-06 evening](#).

The reports move from shipping warnings and avoidance to energy shutdowns, storage constraints and downstream inputs, alongside rerouting and mitigation announcements. Monitor energy availability, freight, insurance, storage and fertilizer together. Estimating a US CPI effect would also require prices, weights, timing, substitution and demand data. Fewer mentions alone do not establish recovery.

Usage, cost and reproduction

The run made **80 paid requests**, including discarded pilots and repairs, using **403,064 input** and **52,964 output tokens**. Applying conservative peak, uncached prices of CNY 2/million input and CNY 8/million output gives **CNY 1.2298**, below the CNY 10 budget. This estimate uses peak uncached prices; actual billing may be lower. See [DeepSeek pricing](#).

The persistent ledger reserves worst-case cost before a call. Ambiguous failures retain their reserve and are not automatically retried. A crash after provider billing but before cache persistence can make an explicit rerun pay again; the ledger still counts both reservations. API credentials are local environment settings and are excluded from the submission.

The submitted notebook reads saved results without internet access or credentials. From the project root:

```
python scripts/check_q4_offline.py
python scripts/build_q4_analysis.py
python scripts/build_q4_timeline.py
python scripts/build_q4_report.py
python scripts/build_notebook.py
```

For a new model run, configure DEEPSEEK_API_KEY locally and use the documented `isw_llm batch` command. Its current archive and cache reproduce this experiment; ongoing collection and a permanent schedule remain separate operational work. See `Q4_System_Design.pdf` for the schema, pseudocode and OOP design, `q4_source_audit.json` for source conservation, and `validation.json` for final software/artifact checks.