

MuseList: Confidence-Gated Multi-Surface Recall for Closed-Book Knowledge Base Construction

Adrian Zimmermann

Technische Universität Dresden, ScaDS.AI Dresden/Leipzig
adrian.zimmermann@mailbox.tu-dresden.de

Abstract

Knowledge base construction requires complete answer sets, not just one plausible fact. We present MUSELIST, a closed-book system for the six-relation AKBC Shared Task 2026. A single open-weight 29.8B-parameter Muse Glimmer model generates candidate answers in several formats, such as recipient lists, recipient–year tables, and venue records. All such records are model-generated; the system retrieves no external documents. Relation-specific rules combine these candidates, select numerical answers using token-level confidence signals, and check proposed reciprocal borders with three separately prompted roles. The complete system scores 0.6700 hidden-test macro-F1 and ranks fourth among twelve systems. Its development validation score is 0.69160; this split informed method selection and is not an untouched evaluation. We describe the method and release an archive that reconstructs the scored submission exactly. Fresh-call support is separate and partial: it re-generates later-stage responses, not an end-to-end submission.

1 Introduction

Language-model knowledge probing commonly asks whether a model can produce or rank a known fact (Petroni et al., 2019; Jiang et al., 2020). LM-KBC instead asks for the complete object set of each subject–relation pair (Alivanistos et al., 2022; Kalo et al., 2025). Correct cardinality may be zero, one, or hundreds: hallucinations lower precision, omissions lower recall, and the best behavior for a country border differs sharply from that for an award with a century of recipients. The 2026 task additionally forbids web retrieval and extra training and caps the sum of inference-time neural parameters at 32B (Kalo et al., 2026).

Our approach separates two problems: eliciting candidate facts and deciding which candidates to retain. Instead of repeating one question, we

ask the same model to produce different structured representations. For an award, for example, one prompt requests a recipient list and another requests a table of recipients and years. We call these prompting formats *surfaces*. They are cues for generation, not retrieved sources or independent factual evidence. Agreement between calls can be useful, but the same model may repeat the same error.

Different relations need different acceptance rules. Long award lists benefit from partitioning the recall task; numerical relations require choosing among conflicting values; land borders admit reciprocal proposals. A prediction that names country *B* as a neighbor of *A* suggests checking whether *B* should also name *A*. City of death retains a direct, conservative prediction route.

MUSELIST combines these relation-specific generation and acceptance steps in one model-based system. Our contributions are the resulting method, an analysis of its cumulative development and final test performance, and a public archive linking model requests and responses to the scored output.

2 Related Work

Prompt wording and answer extraction materially affect factual knowledge exposed by a fixed language model (Jiang et al., 2020; Alivanistos et al., 2022; Ghosh, 2023). Self-consistency samples multiple reasoning paths and aggregates agreement (Wang et al., 2023); our structured-output consensus similarly reuses one model, but varies output format and conditions acceptance on typed records rather than unconstrained prose.

Recent LM-KBC systems make relation-specific policies central. ReWiSe uses relation-wise self-consistency and generated reasoning traces (Albert-Roulhac and Zouaq, 2025). Divide-and-conquer systems separate candidate generation, refinement, and relation-specific

extraction (He and Razniewski, 2025). Other shared-task papers study compact prompting and quantization (Sam, 2025) or reuse one model as generator and judge (Clay et al., 2025). These works motivate explicit call schedules, prompt examples, and relation-level analysis. Our setting differs in two important ways: external retrieval is prohibited in 2026, and all distinct neural components must fit a summed 32B budget.

Our source-form prompts ask the model to complete text resembling RDF, a table, or a URL. The system never opens those URLs or queries a knowledge base. This is structured generation conditioned on the task and permitted training examples, rather than retrieval of external evidence.

3 Task and Constraints

Each input row contains `SubjectEntity`, one of six relations, and an empty `ObjectEntities` list. The relations are award recipient, stock exchange, land border, area, venue capacity, and city of death. We use the post-disambiguation August release: train has 477 rows and validation and test have 475 rows each (validation/test capacity counts are 97/98) (LM-KBC 2026 Organizers, 2026a). Predictions retain exact input order and contain unique strings. The official evaluator computes precision, recall, and F1 per row and macro-averages across rows. It uses maximum bipartite alias matching after Unicode normalization and normalized deduplication; numeric area and capacity predictions match within 5% relative tolerance.

Our only neural component is Muse Glimmer 30B, accessed through OpenRouter. The public weight repository reports approximately 29.8B parameters and an Apache-2.0 license (Meta, 2026; OpenRouter, 2026). DeepInfra, Together, and Fireworks are serving routes for that model, not additional models. Provider fallback is disabled. We use no external retrieval, additional model training, test labels, or manual factual corrections. Released-train labels supply demonstrations and training-side policy selection; exact target keys are excluded from demonstration selection. The appendix records the model route and inference settings.

Why Muse Glimmer? Muse met the parameter and licensing constraints, supported long structured outputs needed for award enumeration, and was available with provider pinning. Released-train probes produced usable JSON, ticker, TSV,

wikitext, and table completions. We therefore fixed this model and studied prompting and aggregation. This was an engineering choice, not a controlled comparison establishing that Muse was the best eligible model.

4 System

4.1 Common inference and parsing

Direct prompts establish initial answer sets. They contain the subject, a relation instruction, a natural-language question, demonstration examples, and an output schema. Stock, borders, and capacity use the first five eligible training rows for that relation; area and city of death use five examples selected for lexical relevance and output diversity. Awards use complementary recall instructions and recipient–year records. The stock route ultimately uses a ticker-based primary-exchange prompt, followed by secondary-listing recall.

The parser extracts JSON or typed records, removes empty markers, and merges duplicate strings after case folding, Unicode decomposition, punctuation removal, and whitespace normalization. It retains the first display form. Numeric relations retain parseable numbers. Final assembly checks the required keys, object strings, and row order before writing the submission.

4.2 Relation policies

An acceptance rule, or *gate*, determines whether generated evidence adds objects, replaces the current answer set, or leaves it unchanged. Table 1 summarizes the rules. Not all require repeated support: some enumeration stages retain every syntactically valid candidate. Rejected selection and verification steps preserve their immediate input, not necessarily the original direct answer.

4.3 Source-diverse set completion

For stock exchanges, a listing-code prompt asks Muse to recall a company’s exchange-qualified ticker internally and return the exchange name. Three further prompts ask for secondary listings using annual-report, market-data, and security-master formats. Only exchanges appearing in at least two of these outputs are appended. These are model-generated views, not documents retrieved from financial sources.

For awards, the system first combines recipient lists and recipient–year records, then asks for additional names through list continuations and alter-

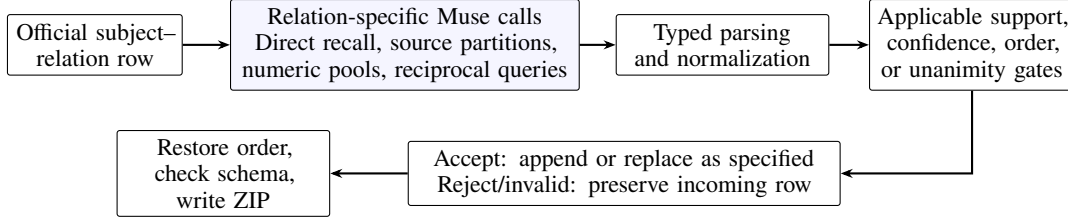


Figure 1: Execution data flow for MUSEList. All blue boxes call the same Muse model; source-shaped strings are never dereferenced. Candidate generation and acceptance are separate, and failure of a gate preserves the incoming complete row. The appendix expands every relation-specific call and acceptance path.

Relation	Candidate source	Acceptance and action	Signal
Award	Lists, tables, and partitioned recall	Combine initial lists; apply the list-table replacement gate; prune aliases and append later support-qualified names	Parsing, length, support
Stock	Ticker-based primary listing; three secondary-listing roles	Generate the primary exchange; append a secondary exchange at normalized support $\geq 2/3$	Cross-role support
Borders	Missing reciprocal graph edge	Append after three unanimous land-border classifications	Unanimity
Area	Twelve numeric representations and an in-pool choice	Cluster candidates at 5%; replace the current answer with a selected value only at margin ≥ 20.7734	Token margin
Capacity	Coordinate surfaces and candidate-order variants	Nonempty surface union replaces the row; subsequent confidence and margin choices replace or preserve it (§4.4)	Parsing, confidence, margin
Death locality	Direct locality or abstention	Normalize one locality conservatively; allow empty	Direct output

Table 1: Final relation policy matrix. Every neural call uses the same Muse model; thresholds were selected on released train.

native table formats. Temporal, geographic, and given-name partitions divide long recall tasks into smaller ones. Further prompts generate award-edition and issuer records associated with candidates. Two groups of checking prompts (“witness families”) require agreement on an award edition and selection of the target award over alternatives, respectively. Their prescribed support rules determine which candidates are added. A separate identity-grouping call, with the award title withheld, removes duplicate name variants within each answer set.

Before identity pruning and the later geographic, issuer, witness, and given-name additions, one award route asks the model to name a Wikipedia-style recipient-list page and then generate its table from memory. It replaces the incoming set only when the generated title has the required list prefix and at least 50 unique recipients parse. The page is never opened. This length rule is not evidence that the generated list is complete or correct.

4.4 Confidence-gated numeric choice

For capacity, prompts first recall venue coordinates and use those generated coordinates to elicit venue-record values. A nonempty union of these surface outputs replaces the preceding capacity set; if empty, that set is retained. Later choice prompts use a broader pool assembled from thirteen numerical sources. Muse must choose a listed value or abstain. We retain a first-pass choice only when the final nonempty-list decision token has log probability at least -0.0001 . A later gate repeats the decision under two deterministic rotations. It counts only in-pool choices with mean token margin at least 22.4922 and accepts the uniquely most-supported value; one qualifying vote suffices, whereas a tie preserves the input. For each token overlapping the rendered answer, the margin is its log probability minus that of the highest-scoring alternative returned in the provider’s token-score list; finite margins are averaged across answer tokens. A rejected later gate preserves its immediate input, including any accepted first-pass choice. For area, twelve distinct prompting formats—for example native units, geometry, gazetteer, temporal and co-

ordinate surfaces—are clustered under the official 5% tolerance. A no-demonstration in-pool choice replaces its input only above mean token margin 20.7734. All thresholds are fixed by strict leave-one-row-out train selection. Token scores are selection signals, not calibrated probabilities of factual correctness.

4.5 Reciprocal-border verification

From the current border predictions we derive only missing reciprocal edges: if A predicts B , both occur as task subjects, and B does not predict A , then $B \rightarrow A$ becomes a candidate. Three fixed Muse roles—a political-boundary editor, physical cartographer, and legal-geography auditor—separately return BORDER or NOT_BORDER at temperature zero. The exact pair is appended only under 3/3 agreement. On released train the gate yields 24 true positives, no false positives, eight true negatives, and no false negatives, including unanimous rejection of the false Qatar–United Arab Emirates proposal. Frozen validation adds South Ossetia to Russia and Palestine to Jordan.

5 Experimental Protocol

Proposed components are checked on released training data before transfer. Candidate prompts, parsers, thresholds, and acceptance criteria are fixed before the corresponding validation or test run. A transformation is applied to the entire split: individual validation rows are not accepted, reverted, or repaired after inspecting their labels. A favorable in-sample slice alone is insufficient; train-neutral or leave-one-out-negative alternatives are rejected.

We nevertheless observed complete-system validation scores repeatedly and used them to select configurations. Validation therefore measures development performance, not an untouched generalization estimate. Table 3 records this cumulative process rather than controlled component ablations.

The development analysis does not use intermediate test submissions as experimental evidence. The final test configuration was fixed before submission. One documented validation/test prompt difference concerns award-list traversal: test uses a train-qualified oldest-first variant, while the reported validation artifact retains the earlier prompt. The acceptance rule is shared; the appendix and archive specify both exact configurations. Cod-

abench evaluates the final 475-row test submission and provides aggregate relation metrics, not private row-level labels.

6 Results

The final system scores 0.6700 hidden-test macro-F1 and ranks fourth among twelve systems in the official leaderboard (LM-KBC 2026 Organizers, 2026b). Table 2 gives its test precision, recall, and F1 alongside development-validation F1. Scores are averaged over rows, not equally over the six relations; award enumeration contributes ten rows despite its long answer lists.

Relation	Test P	Test R	Test F1	Val. F1
Award recipient	.4574	.3345	.3557	.2496
Stock exchange	.8900	.9123	.8400	.8333
Land borders	.9823	.9683	.9652	.9953
Area	.8600	.8500	.8500	.8500
Capacity	.3980	.2449	.2449	.3608
Death locality	.8800	.6100	.5700	.5500
All rows	.7840	.6936	.6700	.6916

Table 2: Final-system test results from Codabench, with development-validation F1 for context. Each split has 475 rows. Validation informed development, and the award traversal prompt differs between splits (§5).

Cumulative development configuration	Val. F1
Initial generation and selection	.66756
Secondary listings + award identity pruning	.67376
Geographic award partitions	.67384
Coordinate-based capacity replacement	.67524
Further source/witness award consensus	.67530
Capacity answer-confidence gate	.68302
Twelve-surface area selection	.68933
Order-varied capacity selection	.69144
Reciprocal-border verification	.69160

Table 3: Cumulative development trace with semantic component names. Values were observed adaptively and are descriptive, not isolated causal effects.

7 Analysis and Discussion

The largest increases in the cumulative validation trace coincide with confidence-based numerical selection. Smaller changes accompany additional set-completion prompts and reciprocal-border verification. These observations help describe development, but cannot establish the isolated contribution or necessity of any component: configurations were added sequentially and chosen using repeatedly observed validation performance. Several additional prompt panels, omission audits, and

numerical aggregation rules failed training-side checks and were excluded.

On hidden test, capacity has the lowest F1 (.2449). One possible failure mode is a confident historical or configuration-specific capacity rather than the target value; aggregate test metrics alone do not diagnose this cause. Award recall (.3345) is lower than precision (.4574), indicating that missing recipients remain a substantial problem; long lists and name variants also create opportunities for false positives. City-of-death recall (.6100) is lower than precision (.8800), consistent with its conservative output policy.

Border prediction is strongest overall, but exact-string reciprocity has a limitation: a proposed reverse edge can name an entity already present under another alias. Inspection of generated test outputs found that six of seven accepted additions were alternate names for already predicted entities. This is an output-identity observation, not a claim about private test labels or a measured test-score contribution.

8 Reproducibility

The public archive provides exact requests, raw responses, deterministic transformations, and intermediate outputs. From a clean checkout, `bash reproduce_improved_submission.sh` OUTPUT reconstructs the submitted 475-row prediction and ZIP byte-for-byte from archived model records. It checks input coverage, schema, order, and artifact identity. Released validation labels permit local scoring; the test score cannot be independently recomputed because its labels are private.

A separate driver can regenerate 2,917 later-stage verification and set-completion responses using recorded inference settings. It excludes the 811 core-generation test calls and does not assemble newly generated responses into a submission. This partial fresh-call facility must not be confused with end-to-end fresh reproduction. Shared prompt templates, split-specific settings, selection rules, and complete row traces are documented in the appendix; the code and archived records are available at <https://github.com/UnGast/muselist-lmkbc2026>.

9 Limitations

Separate prompts to one model can share errors; neither agreement nor high token confidence es-

tablishes factual correctness. Source-shaped outputs can be stale, truncated, or noncanonical. The primary-exchange prompt is narrower than the task’s full listing definition, although secondary-listing prompts seek additional exchanges. The death-city prompt can produce districts or broader localities that conservative normalization does not repair.

Hosted inference does not expose a verifiable immutable serving revision, so new calls need not reproduce archived responses. Exact reconstruction applies to the archived submission, not a guarantee of identical fresh generations. No private test labels were available for row-level error analysis.

10 Ethics and GenAI Disclosure

Generated facts can be wrong and should not be treated as an authoritative knowledge base without verification. The system processes public entity names and challenge files and makes no consequential decisions about individuals. Generative AI assistance was used for experiments, code, analysis, and drafting; all scores, counts, and artifact identities reported here were checked against the archived runs and official evaluator.

11 Conclusion

MUSELIST scores 0.6700 hidden-test macro-F1 using one 29.8B open-weight model and no external retrieval. It combines structured recall formats with relation-specific aggregation and numerical selection rules. Its development trace motivates further controlled comparisons of elicitation and selection. The released archive enables exact reconstruction of the scored submission and inspection of every retained transformation.

References

- Edouard Albert-Roulhac and Amal Zouaq. 2025. [ReWiSe: Relation-wise self-consistency for LLM probing](#). In *Joint Proceedings of KBC-LM and LM-KBC 2025*.
- Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. [Prompting as probing: Using language models for knowledge base construction](#). *arXiv preprint arXiv:2208.11057*.
- Alex Clay, Ernesto Jiménez-Ruiz, and Pranava Madhyastha. 2025. [Judge, generator, executioner: Utilizing an LLM for KBC](#). In *Joint Proceedings of KBC-LM and LM-KBC 2025*.

- Shrestha Ghosh. 2023. [Limits of zero-shot probing on object prediction](#). In *Joint Proceedings of KBC-LM and LM-KBC 2023*, pages 77–85.
- Jingbo He and Simon Razniewski. 2025. [Combining self-retrieval-augmented generation with divide-and-conquer for language model-based knowledge base construction](#). In *Joint Proceedings of KBC-LM and LM-KBC 2025*.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) In *Transactions of the Association for Computational Linguistics*, volume 8, pages 423–438.
- Jan-Christoph Kalo, Tuan-Phong Nguyen, Simon Razniewski, and Bohui Zhang. 2026. [AKBC Shared Task 2026: Predicting complete knowledge base entries from language models](#). In *Proceedings of the 10th Workshop on Automated Knowledge Base Construction (AKBC) at EMNLP 2026*. Association for Computational Linguistics.
- Jan-Christoph Kalo, Simon Razniewski, Bohui Zhang, and Tuan-Phong Nguyen. 2025. [LM-KBC 2025: 4th challenge on knowledge base construction from pre-trained language models](#). In *Joint Proceedings of KBC-LM and LM-KBC 2025*.
- LM-KBC 2026 Organizers. 2026a. AKBC Shared Task 2026: Dataset, relation definitions, evaluator, and baseline. <https://github.com/lm-kbc/dataset2026>. Post-disambiguation August release; accessed 2026-08-12.
- LM-KBC 2026 Organizers. 2026b. AKBC Shared Task 2026: Final leaderboard. <https://www.akbc.ws/2026/shared-task.html>. Final leaderboard linked under Results. Accessed 2026-09-12.
- Meta. 2026. Muse Glimmer 30B model card and weights. <https://huggingface.co/meta-models/Muse-Glimmer-30B>. Accessed 2026-08-12.
- OpenRouter. 2026. Meta: Muse Glimmer 30B model metadata. <https://openrouter.ai/meta/muse-glimmer-30b>. Canonical slug meta/muse-glimmer-30b-20260810; accessed 2026-08-12.
- Fabio Petroni, Tim R  ckt  schel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#) In *Proceedings of EMNLP-IJCNLP*, pages 2463–2473.
- Irene Mary Sam. 2025. [Examining 4-bit quantized Qwen3-8B for object prediction @ LM-KBC 2025](#). In *Joint Proceedings of KBC-LM and LM-KBC 2025*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *International Conference on Learning Representations*.

Appendix: Method and Reproduction Details

Companion to “MuseList: Confidence-Gated Multi-Surface Recall for Closed-Book Knowledge Base Construction”

1 Model Configuration and Evaluation

MuseList uses one approximately 29.8B-parameter open-weight model through OpenRouter: meta/muse-glimmer-30b, canonical slug meta/muse-glimmer-30b-20260810. The public repository <https://huggingface.co/meta-models/Muse-Glimmer-30B> reports 29,776,626,688 BF16 parameters and an Apache-2.0 license. The manifest records the inspected public revision, but the hosted service does not attest that it serves that immutable revision. DeepInfra, Together, and Fireworks are serving routes for this same model; provider fallback is disabled.

The August task release contains 477 training rows and 475 rows each for validation and test. The capacity counts are 97 and 98 respectively on the latter splits. The evaluator computes precision, recall, and F1 per row and then averages over rows. It normalizes strings and matches aliases by maximum bipartite matching; unmatched duplicate aliases can lower precision. Numerical values match within 5% relative tolerance. Empty predictions receive precision 1, and empty gold sets receive recall 1.

All fact-generation calls use only task inputs, permitted training examples, and model-generated intermediate records. No external source is retrieved. Exact subject–relation keys are excluded when selecting demonstrations for a target. The constraints, model-selection rationale, and final test metrics are reported in the main paper. Table 1 retains the more detailed development-validation breakdown.

Relation	Precision	Recall	F1
Award recipient	.3888	.2514	.2496
Stock exchange	.9050	.8850	.8333
Land borders	.9953	.9963	.9953
Area	.8600	.8500	.8500
Capacity	.4897	.3711	.3608
Death locality	.8300	.6000	.5500
All rows	.7970	.7153	.6916

Table 1: Development-validation results for the final configuration. This split was repeatedly observed during method development.

The sections below group operations by relation. Figures show the execution order, and the final traces follow complete rows through all relevant stages. Shared and core prompt templates are reproduced verbatim; the archive contains the exact materialized request for every call, including later-stage prompts. These archived requests, not a shortened prose description, define the precise model inputs.

2 Pipeline Diagrams

Figure 1 expands the paper’s Figure 1 into the actual executable branches. Blue nodes are Muse calls; gray nodes are deterministic. Every call consumes only the subject, relation, frozen instructions, permitted released-train demonstrations, and same-run model outputs shown on its incoming edges. No source-looking string is opened.

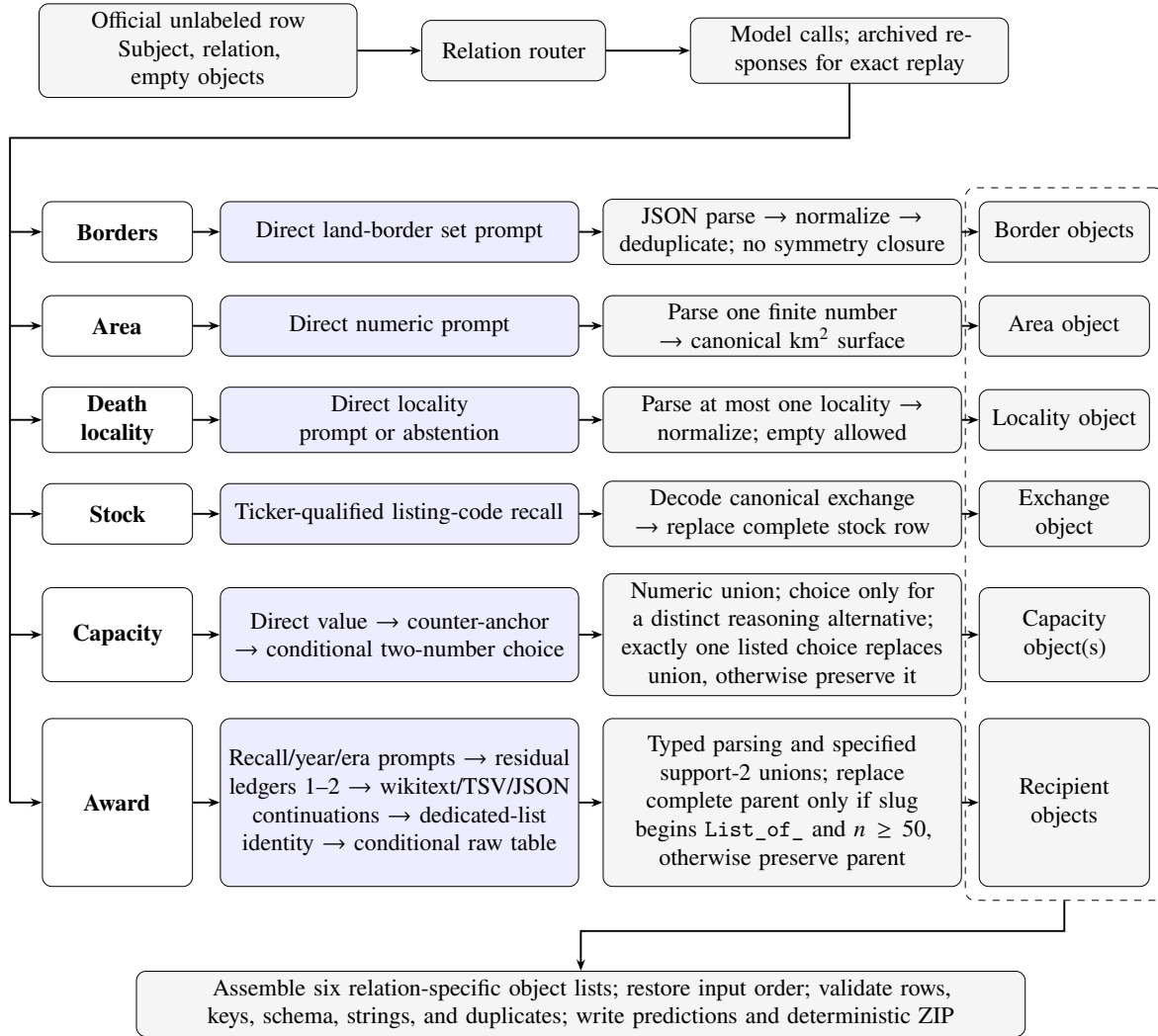


Figure 1: Expanded relation-routing graph. The six horizontal lanes expose the call families, parsers, relation policies, and common assembly. The multi-call capacity and award branches are expanded in Figure 2.

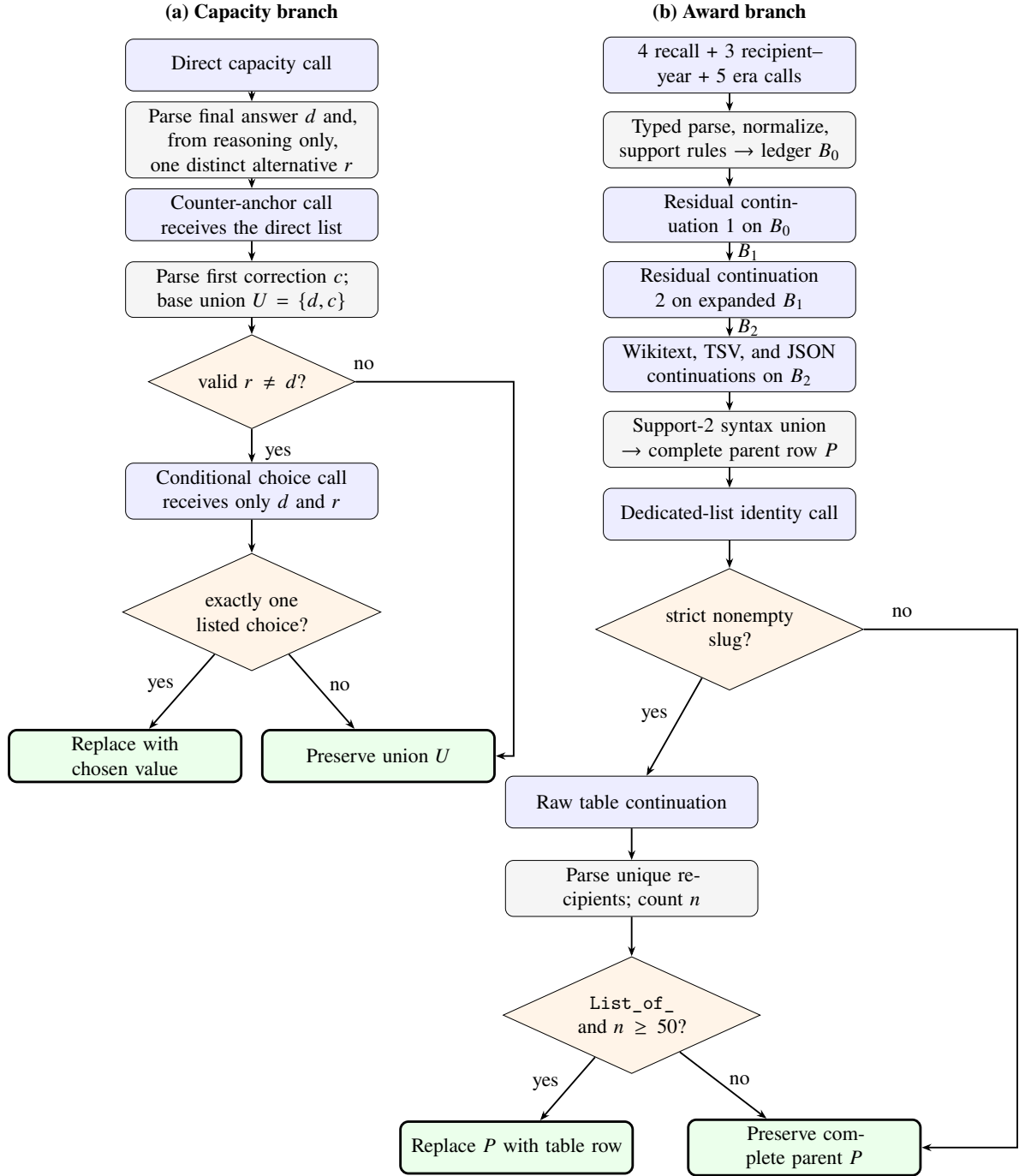


Figure 2: Branch-level expansion for the two multi-call relations. Blue boxes are Muse calls, gray boxes are deterministic transformations, diamonds are executable decisions, and green boxes are the only final outcomes. The capacity choice never receives the counter-anchor value; the award table call exists only after a valid identity result.

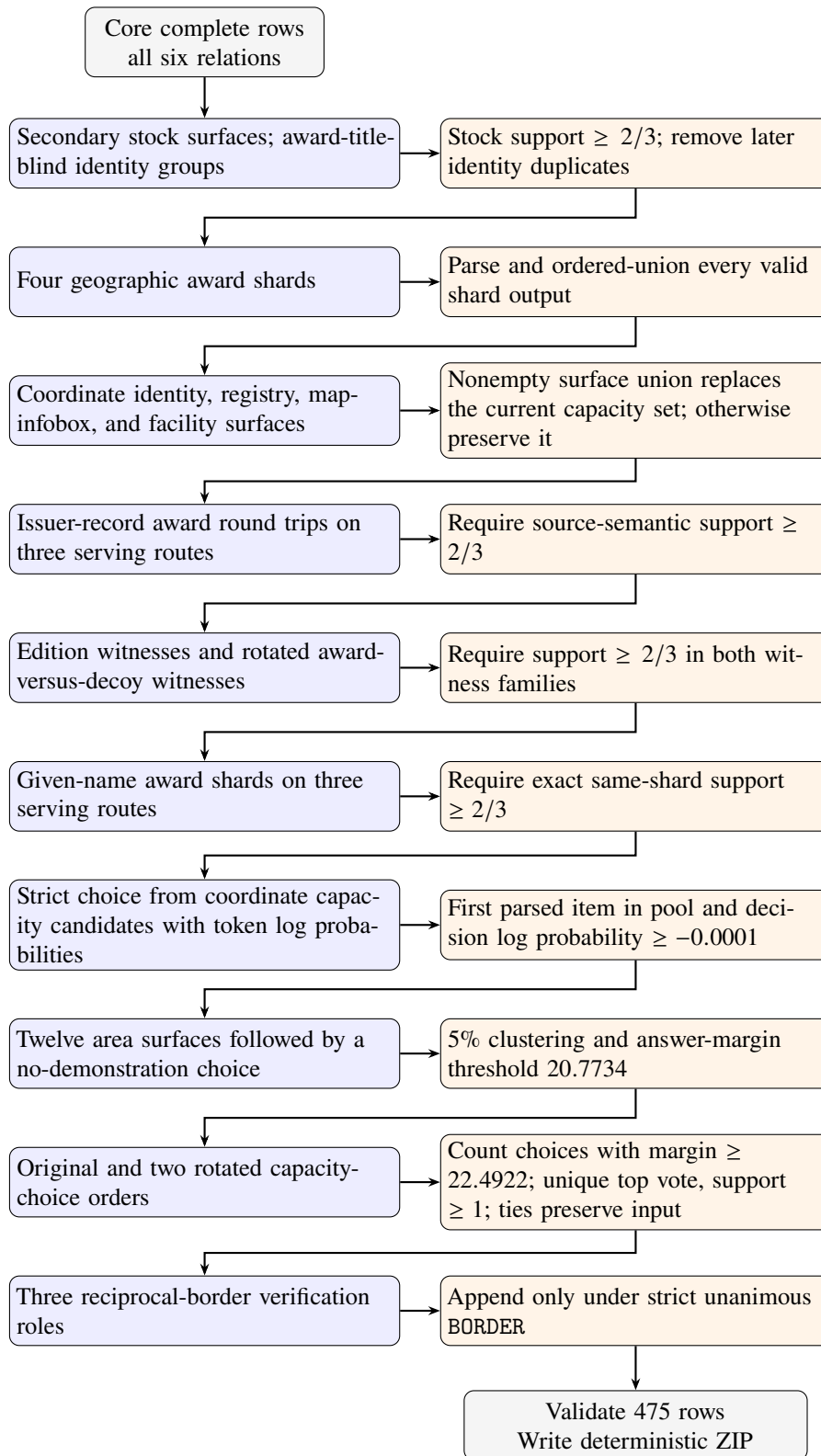


Figure 3: Expanded MUSELIST execution graph. Each blue call family uses Muse Glimmer; each orange box is deterministic and frozen from released train evidence before target transfer. A failed gate preserves its complete incoming row. This figure expands Figure 1 through every final call and acceptance layer.

3 Shared Prompting and Parsing

The shared direct prompt supplies initial candidates for borders, area, capacity, city of death, and an earlier stock route. The final stock route uses ticker-based generation instead. The direct-output examples below illustrate this initial stage only; they are not complete final-system traces.

3.1 Direct prompt and demonstrations

The system message is identical for the five direct relation paths:

```
You are a closed-book knowledge-base
↪ construction system. Use only your
internal knowledge. No web, no retrieval.
↪ Output valid JSON only, with this
exact schema: {"ObjectEntities": ["answer 1",
↪ "answer 2"]}. Use an empty
array when there is no answer. Do not include
↪ explanation.
```

The user message is constructed exactly as follows. Each DEMO is one single-line JSON object copied from released train.

```
Examples:
{DEMO_1}
{DEMO_2}
{DEMO_3}
{DEMO_4}
{DEMO_5}
SubjectEntity: {SUBJECT}
Relation: {RELATION}
Relation definition: {DEFINITION}
Question: {QUESTION}
```

Stock exchange. *Definition:* Return the stock exchange names where this company's shares are publicly traded. If the company is not separately listed, return an empty list.

Question template: Where do shares of {SUBJECT} trade?

Land borders. *Definition:* Return all currently recognized countries or comparable territories that share a land border with the subject. Exclude maritime borders. If there are none, return an empty list.

Question template: Which countries share a land border with {SUBJECT}?

Area. *Definition:* Return one number: total surface area in square kilometers. For countries use total area, including inland water.

Question template: What is the area of {SUBJECT} in square kilometers?

Capacity. *Definition:* Return one integer: the maximum spectator capacity of the venue. Use the highest published capacity if variants exist.

Question template: What is the total capacity of {SUBJECT}?

Death city. *Definition:* Return only the city where the person died. If the person is living or no city is known, return an empty list. Do not return a country or region.

Question template: In which city did {SUBJECT} die?

3.2 Concrete direct examples for every relation

These are materialized test requests, not hand-written illustrations. Stock target TORM receives demonstrations Farfetch → New York Stock Exchange; McColl's → London Stock Exchange; Bangladesh Cement Manufacturers Association → empty; West Japan Railway Company → Tokyo, Nagoya, and Fukuoka Stock Exchanges; and Chubu Electric Power → Tokyo and Nagoya Stock Exchanges.

Border target Somalia receives Sierra Leone → Guinea, Liberia; New Zealand → empty; Germany → its nine listed neighbors; Niger → its seven listed neighbors; and Liberia → Guinea, Ivory Coast, Sierra Leone. No reverse-border fact is added at this direct-generation stage; later verification is separate.

Area target Niutao receives Wellington Island → 5556; Mangareva → 15.4; Israel → 20770; Nantucket → 123.8; and Nigeria → 923768. The requested unit is square kilometers.

Capacity target Pitt Stadium in Pennsylvania receives Sungui Sports Complex → 35000; John A. Farrell Stadium → 7500; Estadio Revolucion Ciudad de Guatemala → 5000; Ibn-e-Qasim Bagh Stadium → 18000; and Nisa-Candybil Stadium → 1500. The archived direct answer is "ObjectEntities": ["78000"].

Death-city target Ad Gerritsen receives Harald Gerlach → Leimen; Marianna Rumjantseva → empty; Peter Eisenberg → Berlin; Hariton Pushwagner → Oslo; and Konrad Rufus Muller → Konigswinter. The prompt explicitly prohibits country or region answers.

The full Unicode-preserving strings and all five complete JSON demonstrations are stored in archived/{split}/packs/nonaward.pack.js on1; the ASCII transliterations above are only for

compact typesetting.

3.3 Demonstration selection

For stock, borders, capacity, and awards, *source* selection takes the first five relation-specific train rows by released-file index after excluding the exact target key. Area and death city use *relevant-diverse* selection. For target t and candidate c ,

$$s(t, c) = .35J(T_t, T_c) + .25J(Y_t, Y_c) \\ + .25J(L_t, L_c) + .15J(G_t, G_c),$$

where T is the set of case-folded word tokens; Y is the subset matching a fixed type vocabulary (e.g., island, city, stadium, university); L is the token suffix after the first “in”, “at”, or “of”; G is the set of character trigrams; and J is Jaccard similarity. Candidates sort by descending score and then source index.

The selector first takes the highest-ranked candidate for each unseen output signature, then fills remaining slots in rank order. Numeric signatures are base-10 magnitude plus object count. Set signatures are object count. This uses labels only to diversify demonstrations, never to inspect a target answer. Selection sidecars record source index, subject, score, and signature for every request.

3.4 Parsing, normalization, and assembly

For every response, code takes the final answer segment, parses the requested JSON, and falls back to conservative list or numeric extraction only when needed. Empty markers are removed. String identity is Unicode decomposition, case-folding, punctuation removal, and whitespace collapse; the first display form is retained. Award support is counted once per frame per normalized name. Capacity numeric identity parses commas and numeric strings and formats the number to 12 significant digits before deduplication.

Rows are then restored to exact official input order and emitted with only the three schema fields `SubjectEntity`, `Relation`, and `ObjectEntities`. Validation checks row count, unique subject–relation keys, exact order, schema, string objects, and duplicate objects. The deterministic ZIP contains exactly one root member, `predictions.jsonl`, with fixed metadata and byte-identical content.

4 Stock Exchanges

The final stock answer combines ticker-based primary-exchange recall with secondary-listing

proposals. The earlier direct/RDF route is described separately below because the replay retains it, not because its answers survive the whole-relation ticker replacement.

4.1 Primary-exchange recall

The ticker surface replaces the entire preceding stock relation. It makes one temperature-zero, 4,096-token call per stock row. Provider fallback is disabled. The system message is verbatim:

```
You recover a company's primary stock
↪ exchange through a closed-book
listing-code memory cue. Use internal model
↪ knowledge only; no tools,
browsing, retrieval, or inference merely from
↪ headquarters. Resolve the
exact company and distinguish an operating
↪ company from similarly named
funds or parents. Return valid JSON only as
{"ObjectEntities":["canonical stock
↪ exchange"]}, or an empty list if the
specific primary listing is not recalled. Do
↪ not return tickers, indices,
countries, or prose.
```

The user template is:

```
Exact company: {SubjectEntity}
Relation: companyTradesAtStockExchange
Memory route: Silently reconstruct the
↪ company's primary quoted ticker
exactly as an investor page writes it,
↪ including an exchange prefix or
suffix (for example EXCHANGE:TICKER or
↪ TICKER.SUFFIX). Decode only that
listing code into the canonical exchange
↪ name. Do not output the ticker.
Return one primary public stock exchange
↪ only. If the company is private,
defunct without a recalled listing, or
↪ ambiguous, abstain.
```

The parser accepts only the requested object list, normalizes exchange names, and writes the component for every stock row. There is no union, fallback, threshold, or validation-selected row mask. The released-train replacement raises stock F1 from 0.7446667 to 0.7550000; validation stock F1 rises from 0.7969048 to 0.8040000 and 21 rows change.

This exact prompt is intentionally preserved for submission reproducibility. It is narrower than the final organizer definition: it requests one primary exchange although multiple current listings and exchange-traded ADR/GDR/CDI venues count. The resulting recall loss is part of the scored method, not a reason to rewrite the archived prompt after evaluation.

4.2 Secondary listings

Three separately prompted roles recall secondary listings from annual-report, market-data, and security-master formats, excluding the generated primary exchange. A normalized exchange name must occur in at least two of the three outputs before being appended. The prompts are not external financial sources; all three responses are generated by Muse. Training-side stock F1 increases from .755 to .771 under this rule, and final validation stock F1 is .83333.

4.3 Earlier direct/RDF route retained in replay

Each stock row receives one additional closed-book source-form call. It uses no demonstrations and does not receive the direct answer. The exact system message is:

You complete a masked corporate source
↪ fragment using only memorized internal model knowledge. Do not browse, retrieve, use
↪ tools, infer from headquarters, or guess. Silently reconstruct the public
↪ stock exchange listing for the exact company, then output valid JSON only as
↪ {"ObjectEntities":["exchange"]}.
Return canonical stock-exchange names, not
↪ ticker symbols, indices, countries, parent-company listings, or prose. Multiple
↪ exchanges are allowed only when independently remembered. Return an empty
↪ list if the company is private, unlisted, a non-listed subsidiary, or the
↪ exchange is not precisely remembered.

The exact user template is:

SubjectEntity: {SUBJECT}
Relation: companyTradesAtStockExchange
Source form: rdf
Replace [MASK] internally, then return only
↪ the JSON answer.

entity label: "{SUBJECT}"@en
wd:COMPANY wdt:P414 wd:[MASK] .

The RDF symbols are a memory-retrieval surface form, not a query: no endpoint, entity identifier, triple, or exchange candidate is supplied. After standard parsing and normalization, assembly applies exactly this row-level rule:

```
if direct_stock_objects is empty and  
↪ rdf_stock_objects is nonempty:  
    final_stock_objects = rdf_stock_objects  
else:  
    final_stock_objects = direct_stock_objects
```

Thus RDF output cannot modify, prune, or add to a nonempty direct stock answer. The rule was selected and frozen on released train before validation inference.

5 Award Recipients

Award completion has three purposes: elicit an initial roster, recover omissions using different output formats and partitions, and apply prescribed support and duplicate-removal rules. Some additions require repeated support; others use a union of all syntactically valid outputs. Those policies are specified separately below.

5.1 Four long-precise frames

Each award first uses the same five source-order award demonstrations and this system message:

You construct closed-book KB entries for
↪ award-recipient lists. The goal is high recall without generic famous-name
↪ hallucination. Output bare recipient entity names only, as JSON. Output valid JSON
↪ only, with this exact schema:
{"ObjectEntities": ["answer 1", "answer 2"]}.
↪ Use an empty array when there is no answer. Do not include explanation.

After the examples, all four user messages contain the exact award, relation definition, and these hard rules:

Hard rules:

1. Include only recipients of this exact
↪ award title.
 2. Do not include nominees, presenters,
↪ jurors, honorary mentions, related awards, predecessor awards, successor
↪ awards, countries, years, or explanations.
 3. Do not pad the list with generally famous
↪ people. If the subject is an honorary degree or membership, famous
↪ activists and politicians are wrong unless you remember this exact institution
↪ awarded them.
 4. Use one canonical name per recipient; no
↪ duplicate aliases, parentheticals, dates, or categories.
 5. If a recipient is a work, album,
↪ organization, or team, output that entity rather than an associated person.
- Question: Which entities have won or received
↪ {SUBJECT}?

The four instruction frames are: (1) recall the official recipient list chronologically and return up to 120 confident names; (2) focus on exact-award winners from authoritative winner-list memory and skip generic celebrities; (3) for honorary degrees or fellowships, require memory of the exact institution; and (4) for prizes or medals, include only exact-award membership. Names are normalized per Section 3.4; a name must occur in at least two of the four outputs.

Concrete example: for test subject Max Planck Medal, the five demonstrations are Nobel Prize in Physiology or Medicine, Premier League Player of the Season, FAI Gold Air Medal, Fulbright Prize, and Ballon d’Or, each with its released recipient list. The first frame begins “Recall the official recipient list chronologically” and asks for up to 120 names. The complete large prompt is preserved in `award_long_precise_4frame.pack.jsonl` rather than abbreviated.

5.2 Recipient-year evidence calls

Three further calls do not see validation/test answers or external facts. They ask the same model to make each proposed recipient carry an internally recalled year. The system message is:

```
You are a closed-book knowledge-base
↳ construction system.
Use only internal model knowledge. Do not use
↳ tools, web search, retrieval,
external corpora, or knowledge-base lookup.
Return valid JSON only with this schema:
{"Candidates":[{"recipient":"canonical
↳ recipient name","year":2000}]}
Use an empty Candidates array when no exact
↳ recipient-year pair is remembered.
Do not explain.
```

The user asks for up to 50 direct recipients of the exact award, paired with a four-digit year, excluding nominees, presenters, committee members, related awards, predecessor/successor awards, uncertain years, and duplicates. The three pass instructions are verbatim:

```
Recall from a chronological official-winners
↳ list.
Recall only recipient-year pairs for which
↳ both fields are specifically
remembered.
Independently reconstruct the precise
↳ recipient-year list and omit doubtful
pairs.
```

Parsing accepts only a nonempty recipient and integer year in 1000–2100. The year is then discarded. All accepted normalized recipient names, i.e. support at least one of three, are unioned with the support-two long-precise result. For Max Planck Medal, the first request ID is 00001-000-awardWonBy-Max Planck Medal-year-anchor. This is evidence conditioning, not year verification against a database.

5.3 Disjoint temporal enumeration

Five additional calls partition the search space rather than repeat the same question. Their system

message is reproduced verbatim:

```
You construct closed-book award-recipient
↳ records using only internal model
knowledge. Do not browse, retrieve, call
↳ tools, use external corpora, or query
a knowledge base. Output valid JSON only as
{"Candidates":[{"recipient":"canonical
↳ name","year":2000}]}
Use an empty Candidates array when no exact
↳ pair is remembered.
```

The user template is:

```
Exact award: {SUBJECT}
Time partition: {ERA}
Enumerate direct recipients whose award year
↳ lies only in this partition. The
time partition is a search cue, not evidence.
↳ Include a recipient only when
both exact-award membership and a four-digit
↳ year are remembered. Exclude
nominees, presenters, jurors, related awards,
↳ predecessor or successor awards,
countries, explanations, and duplicate
↳ aliases. Return up to 80 recipient-year
pairs. JSON only.
```

The five literal substitutions for ERA, in fixed order, are before 1950, from 1950 through 1979, from 1980 through 1999, from 2000 through 2009, and from 2010 through the present. Parsing requires a nonempty recipient and an integer year in 1000–2100. The assembler does not separately re-check the prompt’s era boundary. The year is discarded, every syntactically valid name from the five calls is normalized and unioned, and that component is unioned with the complete preceding award result. No support threshold, era selection, or row fallback is applied.

For test subject Max Planck Medal, the pre-1950 call returns seven pairs, beginning Max Planck (1929), Albert Einstein (1930), and Max von Laue (1931); the other four archived calls return empty candidate lists. Thus only the seven valid names from the first partition can enter this row’s era component. The years are model outputs, not verified or supplied facts.

5.4 Residual list continuations

The first pass starts from the complete award candidate set after stock replacement. The second repeats the same operation from the expanded first-pass candidates. Each split uses ten temperature-zero, 4,096-token calls. The system and user templates are:

```
SYSTEM: Continue the assistant's unfinished
↳ award-recipient archive from
internal memory only. No browsing, tools,
↳ search, explanation, nominees,
```

presenters, jurors, related awards, or
 ↳ placeholders. The bullets already
 written are candidate recipients and must not
 ↳ be repeated. Continue with
 only additional entities that directly
 ↳ received this exact named award.
 Stop when no further distinct recipient is
 ↳ strongly remembered.

USER: Exact award: {SubjectEntity}
 Continue the existing assistant-side ledger
 ↳ at its final empty bullet.
 Do not restart or repeat earlier entries.

ASSISTANT PREFILL:
 == {SubjectEntity}: Recipients ==
 * {existing recipient 1}
 ...
 *

Only text after the final empty bullet is parsed.
 Bullet markers are removed; blank, malformed,
 repeated, and normalized duplicate values are re-
 jected. All distinct residual names are unioned
 with the parent row. The first pass raises released-
 train award F1 from 0.5528621 to 0.5573066 and
 validation award F1 from 0.2052748 to 0.2246002.
 The second raises train F1 to 0.5684942 and vali-
 dation F1 to 0.2318209. No output is selected or
 reverted row by row.

5.5 Agreement across three output formats

This stage makes three temperature-zero, 4,096-
 token calls per award row with minimal reasoning
 effort. All variants share this system message:

Continue the assistant's unfinished award
 ↳ archive from internal memory
 only. No browsing, tools, search,
 ↳ explanation, nominees, presenters,
 jurors, related awards, or invented
 ↳ placeholders. Resolve the exact named
 award and emit only entities that directly
 ↳ received it. Recipients may be
 people, organizations, teams, or works.
 ↳ Continue the supplied syntax
 rather than restarting it.

The three verbatim user/prefill forms are:

WIKITEXT USER: Exact award: {SubjectEntity}
 Relation: awardWonBy
 Continue the unfinished recipients section
 ↳ with one direct recipient per
 bullet. Do not emit years.
 PREFILL: == {SubjectEntity}: Recipients ==
 *

TSV USER: Exact award: {SubjectEntity}
 Relation: awardWonBy
 Continue the unfinished archival table with
 ↳ one award year and direct
 recipient per row, tab-separated. Use the
 ↳ exact award's complete remembered
 history.

PREFILL: award<TAB>year<TAB>recipient
 {SubjectEntity}<TAB>

JSON USER: Exact award: {SubjectEntity}
 Relation: awardWonBy
 Continue the unfinished JSON array with
 ↳ direct recipient names only.
 Produce valid JSON and close the array.
 PREFILL: {"award": "{SubjectEntity}",
 ↳ "recipients": ["

Each syntax is parsed independently. A normal-
 ized name contributes at most one vote per syntax;
 names with support at least two are unioned with
 the second residual-pass parent. The threshold was
 selected on released train, where award F1 rises
 from 0.5684942 to 0.5700759. Validation award
 F1 rises from 0.2318209 to 0.2319206.

5.6 Recipient-list identity and table generation

This roster-generation stage adds no non-award
 calls. Identity uses temperature 0 with provider
 fallback disabled and this exact prompt pair. The
 test call ceiling is 768 tokens; validation uses 1,536
 after the same frozen prompt was rerun to cor-
 rect reasoning-budget exhaustion. No top-p or
 reasoning-effort override is sent.

SYSTEM: Resolve the exact English Wikipedia
 ↳ article whose primary content
 is the chronological list, table, or roll of
 ↳ direct recipients of this
 exact award. Work from internal memory only
 ↳ and do not name recipients.
 Return JSON only:
 ↳ {"list_article_slug":string|null,
 "table_section":string|null}. Use the exact
 ↳ title slug after /wiki/ with
 underscores. Do not return the general award
 ↳ article unless its primary page
 itself contains the recipient table. If no
 ↳ exact list surface is specifically
 remembered, use null. No browsing, tools,
 ↳ guessed titles, placeholders, or
 prose.
 USER: Exact award: {SubjectEntity}

For a strict nonempty slug, ledger continuation
 uses 8,192 tokens and:

SYSTEM: Continue the unfinished raw-wikitable
 ↳ recipient roll from internal
 pretrained memory of the identified list
 ↳ article. Emit chronological rows
 only as `year || direct recipient`, one
 ↳ recipient per row. Continue as many
 specifically remembered rows as possible.
 ↳ Exclude nominees, teams unless
 directly awarded, related awards, prose,
 ↳ headings, guesses, browsing, and
 tools.
 USER: Continue only table rows.
 ASSISTANT PREFILL:

```

SOURCE: https://en.wikipedia.org/wiki/{slug}
RAW WIKITABLE EXPORT FOR EXACT AWARD:
↪ {SubjectEntity}
|-
| year || recipient
|-
|

```

This is the deterministic traversal used in the final system. On the test split, the frozen independently train-positive forward traversal is used. Its system message says “Begin with the oldest specifically remembered edition and work forward.” Its assistant prefix places ORDER: CHRONOLOGICAL (OLDEST TO NEWEST) immediately after the RAW WIKITABLE EXPORT FOR EXACT AWARD: heading and before the table rows. That variant improved released-train award F1 to 0.5732859, was frozen, and was run on validation before test; its validation run was a no-op because Turing Award yielded 46 parsed names, below the unchanged 50-name gate. Thus the gate and source-surface method are shared, but the two scored split artifacts do not use a byte-identical ledger prompt. Both exact prompt packs are archived and the fresh driver pins the split-specific choice explicitly.

The URL-shaped string is never dereferenced. Slug parsing requires one JSON object and a conservative character grammar. Ledger parsing accepts only well-formed rows with integer year 1000–2100 and nonempty recipient. The parent row is replaced iff the slug begins `List_of_` and at least 50 unique normalized recipients parse. Otherwise the complete multi-syntax row survives. These predicates are target-label-independent after train-side selection. The 50-name cutoff is a substantial-output gate; it does not prove that the list is complete, factually correct, or longer than its parent.

Examples of replacement and preservation.

For validation Turing Award, identity returns `List_of_Turing_Award_laureates`. Strict ledger parsing yields 78 unique recipients. Both gate predicates pass, so the 99-name multi-syntax row is replaced. For Presidential Medal of Freedom, identity returns `List_of_Presidential_Medal_of_Freedom_recipients`, but only nine names parse; the parent count of 120 is preserved. For test Nobel Prize in Literature, identity returns `List_of_Nobel_laureates_in_Literature`, 119 names parse, and the 127-name parent is replaced. Thus success, short-ledger fallback, and null/invalid-slug fallback are all executable paths.

5.7 Identity grouping and partitioned recall

Award identity pruning. *Candidate construction:* Award-title-blind within-row identity grouping

Frozen acceptance and effect: Keep the lowest-index surface in every duplicate group

Geographic award shards. *Candidate construction:* Europe; Americas; Asia/Oceania; Africa/stateless/multinational

Frozen acceptance and effect: Support-one union of every valid parsed name from the four disjoint shards

Given-name award shards. *Candidate construction:* The same first-token shard through three pinned serving routes

Frozen acceptance and effect: Add only exact same-shard support $\geq 2/3$

5.8 Issuer-record and two-family checks

Issuer-record checks ask the model to associate a candidate with an award edition and a source identifier, then reconstruct the recipient. A response qualifies only if its verification field is true, its reverse-recipient name normalizes to the candidate, and its edition/source fields pass the fixed issuer-source predicate. At least two of three serving-route responses must qualify. This is consistency checking of generated fields, not opening or verifying an external source. Conservative identity matching prevents adding a name already represented in the row.

The two-family check uses three edition prompts and three reverse award-choice prompts. The edition family requires at least two `received=true` responses agreeing on a parseable edition. The reverse family requires at least two choices of the target award among the supplied alternatives. A candidate is added only if both families pass and it is not an existing identity. These separate prompt families may still share model errors.

6 Venue Capacity

The initial capacity route forms a numerical answer set and may reduce it to one value. Coordinate-conditioned generation can then replace that set. Later confidence-based choices operate on a broader candidate pool. Keeping these stages separate is important: rejection preserves the immediate input, which may differ from the original direct answer.

6.1 Correction of the initial draft

The direct capacity list is passed as an explicitly untrusted draft to one more call of the same model. The system message is:

You perform closed-book factual error
→ correction using only internal model knowledge. Do not browse, retrieve, call
→ tools, or estimate. Output valid JSON only as {"ObjectEntities":["candidate
→ 1","candidate 2"]}. Return at most three distinct numeric candidates in
→ descending confidence, or an empty list if no exact value is remembered.

The user template is:

SubjectEntity: {SUBJECT}
Relation: hasCapacity
Untrusted draft candidates: {BASE_JSON}
Audit frame: Treat the draft values as
→ potentially misleading anchors.
Independently recall the exact published
→ fact, then correct the draft.
Return normal spectator/event capacity as a
→ bare number, not pitch area, attendance, or temporary expansion.
The draft is evidence of what another attempt
→ guessed, not a fact. Return only corrected JSON.

Only the first parsed correction value is retained. It is unioned with the base under numeric identity; no value is multiplied, rounded, range-expanded, or selected using validation labels. Concrete example: Pitt Stadium's base draft is ["78000"]; the correction call returns "ObjectEntities":["78000"], so the final set remains one value. A changed correction would produce two candidates, trading precision for recall.

6.2 Choice between direct and reasoning candidates

Capacity self-adjudication adds one same-model call only when the saved direct response exposes two distinct candidates. Candidate extraction is deterministic. From the final answer segment, parse the first numeric value as direct. From the provider-returned reasoning, discard everything before the final literal Question: marker so numbers in demonstrations cannot become candidates. Recognize positive values from 100 through 500,000 using either 3–6 digits or comma-grouped digits. Exclude numeric identity equal to direct. Rank remaining values by descending mention count and then descending final mention position. Retain only the first as alternative. A row

is eligible if direct and alternative are both present and distinct.

The adjudication system message is exactly:

You construct closed-book numeric KB entries
→ for venues. Choose only from the provided candidate capacity numbers. Do not
→ use web, retrieval, databases, tools, or external lookup. Output valid JSON
→ only with schema {"ObjectEntities": ["candidate"]}. Use an
→ empty array if none of the candidates is likely to be the official or
→ commonly listed capacity.

The user template is:

SubjectEntity: {SUBJECT}
Relation: hasCapacity
Definition: maximum spectator or event
→ capacity of the exact venue/facility.
Candidate capacity numbers from the current
→ system: {CANDIDATE_JSON}
Task:
1. Identify the exact venue named by the
→ subject, using location text only for disambiguation.
2. Select at most one candidate number that
→ is likely to be the official or commonly listed capacity.
3. Return [] if the candidates look like
→ generic guesses, city populations, years, areas, or if you are unsure.
4. Do not return any number that is not in
→ the candidate list.

There is one call per eligible row at temperature 0, top-p 1, low reasoning effort, and 512 maximum tokens. Parsing normalizes the returned values and matches them against the two listed values. Exactly one distinct listed match becomes the adjudication component. Zero matches, an empty response, an unlisted value, or two listed values is an abstention. The frozen *fallback-nonempty* rule is:

```
if adjudication_component contains exactly  
→ one listed candidate:  
    final_capacity_objects =  
        → adjudication_component  
else:  
    final_capacity_objects =  
        → pre_adjudication_capacity_objects
```

The call does not receive the counter-anchor output, validation gold, a retrieved value, or an external candidate vocabulary. Both candidates were already generated inside the same direct Muse call.

For validation subject St Andrew's in Birmingham, the direct answer is 29409 and the top distinct reasoning mention is 30079. The pre-adjudication union contains both numbers; the frozen choice call returns only 30079. The released validation value is 30016, so both old

numbers fall within the evaluator’s 5% tolerance but the second false-positive slot lowers precision. Selecting one value improves this row. This example was not used to define the rule: on all released-train capacity rows, the policy changed 20 rows and raised F1 from 0.23667 to 0.25667 before validation inference. The frozen validation transfer made 70 calls, changed six complete rows, and raised capacity F1 from 0.30241 to 0.30584. Test has 67 eligible rows and ten replacements. Gold answers and row-level correctness remain private; Codabench reports aggregate capacity precision/recall/F1 .3980/.2449/.2449 for the final submission.

6.3 Coordinate-conditioned generation

The model first generates coordinate information and then recalls capacity values in venue-record formats. The surface-union transformation parses at most the first value from each response variant, normalizes and deduplicates those values, and replaces the current row if this union is nonempty. Otherwise it retains the current row. This is a prediction transformation, not merely construction of an unused candidate pool.

The later choice prompts use a broader thirteen-source numerical pool, combining the specified capacity sources. Pool generation and row replacement are distinct operations. The complete source packs and their ordering are recorded in the archive.

6.4 Confidence and candidate-order selection

Capacity confidence. *Candidate construction:* The prompt requests one choice from the supplied coordinate pool. The parser takes the first parsed item, requires it to be in the pool, and ignores any additional items. This parser is also used for area and order-varied capacity selection.

Frozen acceptance and effect: Replace with that one item only when the final nonempty-list decision token has log probability ≥ -0.0001

Order-varied capacity. *Candidate construction:* Original candidate order and two deterministic rotations

Frozen acceptance and effect: Count only in-pool choices with margin ≥ 22.4921875 across original/rotated orders. Accept the uniquely most-supported value with support ≥ 1 ; a tie or no qualifying vote preserves the input. One qualifying vote suffices; cross-order agreement is not required.

For each token overlapping the selected answer, the margin is the chosen token’s log probability minus the largest alternative log probability returned in the provider’s top-token list. Only finite margins are averaged. A missing usable feature cannot pass the corresponding threshold. The selectors use Fireworks with temperature zero, provider fallback disabled, and completion log probabilities enabled. A confident model answer can still be factually wrong.

7 Geographic Area

Twelve-format area. *Candidate construction:* Temporal, geometry, gazetteer/prefill, Wikipedia-like, native-unit, coordinate, and core numeric surfaces

Frozen acceptance and effect: Cluster at 5%; use a strict in-pool, no-demonstration choice only at $\text{answer_mean_margin} \geq 20.77343735098843$

The twelve representations are distinct prompting formats, not statistically independent measurements. Numerical candidates are generated from recalled temporal, geometric, gazetteer, Wikipedia-like, native-unit, coordinate, and direct-answer records. The model selects a listed candidate without demonstrations; the fixed answer-margin rule determines whether it replaces the incoming answer. A rejected choice preserves that answer. The same token-margin definition and pinned serving configuration described for capacity apply here. Exact prompt texts and candidate ordering are preserved in the request packs.

8 Land Borders and City of Death

Borders begin with the shared direct set prompt. For every generated edge $A \rightarrow B$, a reverse proposal $B \rightarrow A$ is eligible only when both countries are task subjects and that exact reverse string is absent. Three fixed Muse roles—political-boundary editor, physical cartographer, and legal-geography auditor—return strict BORDER or NOT_BORDER decisions. Only three BORDER decisions append the proposed neighbor; otherwise the row is preserved. The calls use Fireworks, temperature zero, and disabled provider fallback. Exact-string absence is not entity-level absence: an accepted name can duplicate an existing entity under a different alias.

City of death uses only the shared direct locality-or-abstain prompt and conservative parsing. No additional completion or verification layer changes

this relation. A returned district or broader locality is not automatically converted into the correct city.

9 Development Protocol and Call Accounting

Each deployed policy was specified and checked on released train before its target-split execution. Target rows were excluded from demonstrations where relevant, and no validation row was accepted, reverted, or repaired individually. Complete-system validation was observed repeatedly, so the main paper’s cumulative trace documents development rather than estimating isolated causal effects.

Train-neutral or leave-one-out-negative alternatives were not transferred. Rejected experiments are not required by, and do not enter, the final replay. Fine-grained train diagnostics remain available in the public run reports but are omitted here because they are experiment history rather than part of the method specification.

9.1 Training-side evidence

Secondary stock support improves released-train stock F1 from .755 to .771. The first capacity confidence threshold improves the 69 coordinate-matched rows under strict leave-one-row-out evaluation from .25604 to .26087 and is selected in 68/69 folds. The area margin rule improves strict leave-one-row-out area F1 from .81 to .82 and is selected in 99/100 folds. The order-varied capacity rule reaches .27536 in leave-one-row-out evaluation and is selected in 68/69 folds. The border verifier produces 24 true positives, no false positives, eight true negatives, and no false negatives on released train. Neither validation gold nor evaluator output appears in a model payload.

Validation scores were nevertheless observed after complete frozen transformations, so development is adaptive. Source-consensus components are associated with the first cumulative validation change; numeric confidence and order gates with the largest later changes; and reciprocal verification with the final .691599 value. This trajectory is development evidence, not a held-out estimate or an estimate of isolated causal effects.

9.2 Core inference settings

Core direct and four-frame award calls use temperature 0, top-p 1, reasoning effort low, and maximum 4,096 non-award or 8,192 award tokens.

Award-year calls use temperature 0.35, top-p 0.95, and 8,192 tokens. Counter-anchor calls use temperature 0, top-p 1, and a 4,096-token API ceiling. Award-era calls use temperature 0.2, top-p 1, reasoning effort low, and an 8,192-token API ceiling. Stock RDF calls use temperature 0, top-p 1, reasoning effort low, and a 512-token API ceiling. Each request has a 360-second timeout and two retries; failed coverage is reconciled and retried once. Assembly requires exactly one response for every materialized request. Capacity self-adjudication also uses temperature 0, top-p 1, reasoning effort low, and a 512-token ceiling.

9.3 Call coverage and runtime

The retained core rows use 817 validation and 811 test calls: awards 185/181, stock 100/100, borders 68/67, area 100/100, capacity 264/263, and city of death 100/100. The earlier direct-generation stage instead makes 852/850 calls, including 200 stock calls later superseded by 100 ticker-based calls and only 120 initial award calls before continuations. These totals refer to different stages; neither is a whole-system cost estimate.

	Split	Calls	Tries	Sum (s)	Max (s)
Val.	817	847	4,753.005	116.198	
Test	811	840	4,397.442	34.608	

These retained-core figures are an archive audit, preserved as `execution_accounting.json`; they are not regenerated by the prediction replay’s final PASS assertion. The latency column sums per-request elapsed times and is not wall-clock runtime because calls were concurrent. Archived response objects do not contain provider billing token counts or charged amounts, so an exact historical cost cannot be reconstructed without the private account ledger. Token ceilings alone do not determine actual billing.

The verification and set-completion layers comprise 2,917 successful test calls: 310 for secondary listings and identity pruning, 40 geographic award shards, 71 coordinate capacity surfaces, 143 source-semantic award checks, 786 dual-witness award checks, 120 given-name checks, 17 capacity-confidence choices, 1,375 area surface and selector calls, 34 rotated capacity choices, and 21 reciprocal border checks. DeepInfra, Together, and Fireworks are serving routes for the same Muse model. Response metadata confirms Muse for every archived call; an obsolete exporter label in one

area request pack is overridden by the public fresh-call driver and did not determine the served model.

10 Complete Final-System Row Traces

All six examples start at the official test input: a subject, a relation, and an empty `ObjectEntities` placeholder, not an intermediate prediction. They trace archived model outputs and frozen transformations; none asserts correctness against hidden test labels. Source-shaped prompts are closed-book model calls, not retrieved records. Components targeting other relations leave the row unchanged. Complete ordered arrays are in `dataset2026/reproduction/muse_glimmer30b_improved_candidate/checkpoints/test/10_reciprocal_border_verification.jsonl`; the line numbers below refer to that file and the official test input. The companion `TRACE_EVIDENCE.md` locates every preceding stage.

10.1 Award recipients: Sydney Peace Prize

The input is Sydney Peace Prize, `awardWonBy` (line 205). All four long-precise recall frames and all three recipient-year calls parse empty. Of five temporal calls, only the final era supplies a parsed name, Adolfo Pérez Esquivel; its generated year is not a gold fact. The first residual continuation adds nothing; the second appends Mairead Maguire and Malala Yousafzai to this one-name row. Three-syntax consensus adds nothing: the wikitext and TSV lists have no common name, and the malformed JSON continuation parses as fragments rather than its apparent recipient list. The dedicated-list identity response supplies no qualifying slug, so no table replacement occurs and the three-name row survives.

Within-row identity grouping returns no duplicate groups. Geographic shards then add Desmond Tutu from the Africa/stateless shard under support-one union; the other three shards add nothing. Source-semantic consensus yields no qualifying addition for this award. The dual-witness gate also adds nothing: for example, the candidate The International Campaign to Abolish Nuclear Weapons passes edition consensus for the generated year 2017 but has only one supporting reverse-title witness, below two. Given-name shard consensus adds nothing. Later numeric and border stages preserve the entire ordered four-name output:

[Adolfo Pérez Esquivel, Mairead

Maguire, Malala Yousafzai, Desmond Tutu].

This is the complete emitted row, not merely the last additions or a claim that these four names exhaust the real award's recipients.

10.2 Stock exchange: TORM

The input is TORM, `companyTradesAtStockExchange` (line 209). The five-demonstration direct answer parses to [Nasdaq Copenhagen]. An archived RDF-shaped answer also gives Nasdaq Copenhagen, but its empty-base-only fallback cannot change this nonempty row. Whole-relation ticker-surface replacement independently returns the same singleton. The three secondary-listing prompts exclude Nasdaq Copenhagen: annual-report returns an empty list, while market-data and security-master each return New York Stock Exchange. Normalized support is therefore exactly 2/3, meeting the frozen append threshold. No later component targets stock; the complete final output is [Nasdaq Copenhagen, New York Stock Exchange].

10.3 Land borders: Zambia

The input is Zambia, `countryLandBordersCountry` (line 373). The five-demonstration direct call yields the ordered list [Angola, Botswana, Democratic Republic of the Congo, Malawi, Namibia, Tanzania, Zimbabwe]. Parsing retains all seven surfaces, and every stage before reciprocal verification preserves this row. The generated row for Mozambique (line 311) already contains Zambia; both countries are input subjects, and the reverse surface is missing from Zambia's row. This generates the candidate Zambia–Mozambique without consulting labels. Political-boundary editor, physical cartographer, and legal-geography auditor each return the strict JSON decision BORDER. Unanimous 3/3 acceptance appends Mozambique, giving the complete final array [Angola, Botswana, Democratic Republic of the Congo, Malawi, Namibia, Tanzania, Zimbabwe, Mozambique].

10.4 Area: Niutao

The input is Niutao, `hasArea` (line 1). The target-excluded five-demonstration direct call returns ["2.5"]; all stages before the area selector preserve it. In the twelve-source pool, temporal and

shoreline components also supply 2.5; bounding-box reconstruction supplies 2.07; survey-ledger, infobox prefill, and coordinate-area components supply 2.99. Gazetteer, Wikipedia-like, and all three native-unit components contribute no parsed value; the core row supplies 2.5. The frozen 5% clustering therefore gives the complete ordered pool ["2.5", "2.07", "2.99"]. The no-demonstration, in-pool choice returns ["2.99"] with `answer_mean_margin` 21.5625, exceeding 20.77343735098843. It replaces 2.5. Subsequent capacity and border gates do not target this relation, leaving the final array ["2.99"] in square kilometres. The two unselected numeric alternatives are not unioned into the final answer.

10.5 Capacity: Pitt Stadium in Pennsylvania

The input is Pitt Stadium in Pennsylvania, `hasCapacity` (line 101). Direct recall returns ["78000"]; the counter-anchor call repeats 78000, so their numeric union remains a singleton. The direct reasoning also mentions 56000. The conditional self-chooser receives ["78000", "56000"] and selects 78000, preserving the core row. Later coordinate identity returns (40.4444, -79.9569); registry, map-linked infobox, and geocoded-facility calls each return 56000. Their support-one surface union is ["56000"] and *replaces*, rather than extends, the earlier row.

The separate thirteen-source selector pool is ["78000", "56000", "35000", "1925", "1999", "10000", "5000", "20000", "10300"]. It includes reasoning-derived and prior candidates, not just the three coordinate answers; even the year-like values are preserved as archived. The coordinate-conditioned chooser returns 78000 with decision-token log probability approximately $-4.76837 \times 10^{-7} \geq -0.0001$, so the confidence gate changes 56000 back to 78000. Original order and two rotations then select 78000, 56000, and 78000, with answer margins 22.78125, 23.1875, and 23.1875. All exceed 22.4921875: 78000 wins by two qualifying votes to one and the order-varied gate leaves the row unchanged. Border verification does not target capacity. Final output: ["78000"].

10.6 Death locality: Ad Gerritsen

The input is Ad Gerritsen, `personHasCityOfDeath` (line 376). The target-excluded selector supplies demonstrations

for Harald Gerlach, Marianna Rumjantseva, Peter Eisenberg, Hariton Pushwagner, and Konrad Rufus Müller. The direct city-or-abstain response is {"ObjectEntities": ["Amsterdam"]}. Parsing retains this single locality; normalization changes neither its spelling nor cardinality. There is no death-specific secondary gate in the final pipeline. Every subsequent checkpoint preserves it, yielding the complete final array ["Amsterdam"]. This is a test prediction, not a gold-verified death-place assertion.

11 Reproducing the Submission

11.1 Exact final-system rebuild

From a clean checkout, one command first bootstraps the four pinned Python packages if needed, then executes the complete final pipeline from archived model records and verified deterministic components, validates the official test schema and order, and creates the exact submitted ZIP:

```
bash reproduce_improved_submission.sh
↪ /tmp/muselist-improved-replay
```

An independent clean temporary-directory replay produced the submitted 475-row prediction file and one-member ZIP byte-for-byte. The command exits nonzero on any artifact, schema, order, or package mismatch. The machine-readable manifest maps every semantic component to its hand-off, prompt pack, raw responses, input, deterministic transformer, output, and evaluator report. Codabench submission 890031 evaluated this frozen package and reports displayed macro-F1 0.6700. The score is release metadata: local replay proves artifact identity but cannot recompute hidden-label evaluation.

11.2 Partial fresh-response regeneration

The independent fresh-response audit is invoked from the repository root with:

```
export LMKBC_API_BASE=https://openrouter.ai/api/v1
export LMKBC_API_KEY=...
python3 dataset2026/reproduction/\
  muse_glimmer30b_improved_candidate/rerun_model_calls.py
↪ \
--output-dir /tmp/muselist-fresh-calls
↪ --resume
```

The driver reads the public call manifest and executes every verification and set-completion prompt pack with a global Muse-model override, explicit provider order, disabled fallback, temperature zero,

recorded token ceiling, and the recorded reasoning and log-probability options. The 811 core-generation test calls are not part of this fresh-response driver. Credentials are accepted only through environment variables and are not written to the repository. A `--plan-only` mode prints the complete 2,917-call schedule without network access.

The frozen request packs contain no target labels or evaluator output. This mode regenerates fresh responses and tests whether the same prompts remain stable under a new hosted-model execution; it does not yet assemble those new responses into a fresh prediction and does not promise the historical response bytes or validation score. Exact artifact reproduction uses the archived command above.

The response writer records request IDs, returned provider/model metadata, attempt counts, elapsed time, reasoning text when supplied, completion log-probabilities when requested, and usage fields when the provider supplies them. Resume skips only already recorded request IDs. A fast contract test plus the complete archived-response regression is available as:

```
bash test_reproduction_contract.sh
```

The full regression executes every archived component, then checks the final prediction and ZIP identities. It does not turn a fresh provider rerun into an exact reproduction: new responses, validation score, and test ZIP may differ.

11.3 Artifact navigation

The repository at <https://github.com/UnGasT/muselist-lmkbc2026> contains the reproduction root under `dataset2026/reproduction/`. Canonical request packs specify system and user messages, assistant prefixes, model/provider settings, token limits, request IDs, and dependencies. The core archive contains shared prompts and responses; the final-method bundle contains later-stage packs, responses, policy provenance, and independently stored stage outputs. The entry-point script resolves both automatically.

The released code and bundled task material retain their licenses and attributions in the repository's NOTICE. Model responses are distributed as experiment records. Credentials for optional fresh calls are supplied through environment variables, not stored in the archive.