

Advisor Learning Data Processing Project

1. Source Material

- **Input Files**

- `advisor_learningOBJ.jsonl` – structured logs of advisor conversations (`Advisor_1`, `Advisor_2`, `Advisor_3`) with roles (`User`, `Kimi`), turns, and verbatim messages.
- Derived working files: `advisor_learning_batch_1.json`, `advisor_learning_batch_2.json`, `advisor_learning_batch_3.json`, and `advisor_learning_first_semantics.csv`.
- **Objective:** Extract, normalize, and enrich conversational data into machine-readable learning fragments representing advisor reasoning.

2. Sorting and Canonicalization

- Ordered by Advisor (1→3), Turn (ascending), Role (User before Kimi).
- Assigned ID prefixes: Advisor 1 = 1XXX, Advisor 2 = 2XXX, Advisor 3 = 3XXX.
- Each record now precisely traceable to its origin.

3. Learning-First Semantic Extraction

- **Goal:** Derive cognitively meaningful fragments (learning-oriented statements) from free-form text.
- **Method:** Rule-based detection of learning verbs (`define`, `clarify`, `design`, `reflect`, `verify`, `ensure`, `check`, `evaluate`, `plan`, `learn`, etc.).
- **Segmentation Logic:** Split sentences by punctuation, filter those with learning verbs, further split by conjunctions, discard fragments shorter than four words.
- **Outcome:** 3,248 learning-first fragments extracted.
- **Output:** `advisor_learning_first_semantics.csv`

4. Data Structuring

- **File:** `advisor_learning_first_semantics.csv`
- **Columns:** `id`, `user`, `turn`, `source`, `learning_first_semantic`
- **Characteristics:** Clean dataset, ordered, ID-traceable, each representing one cognitive action or thought.

5. Weighting and Rationale Modeling (Planned)

- **Objective:** Quantify relevance of each fragment to advisor reasoning.
- **Design:** Assign weights 0.1–0.99 and a rationale for each.
- **Algorithmic Heuristics:**
 - Reflective (0.80–0.99)
 - Procedural (0.55–0.79)
 - Descriptive (0.30–0.54)
 - Minimal (0.10–0.29)
- **Rationale Templates:** Explain classification using reflection or process-based language.
- **Output:** `advisor_learning_weights.csv`

6. Execution Summary

- Sorting, semantic extraction, and validation completed.
- Pending: Weight assignment, validation, consolidation for training dataset.

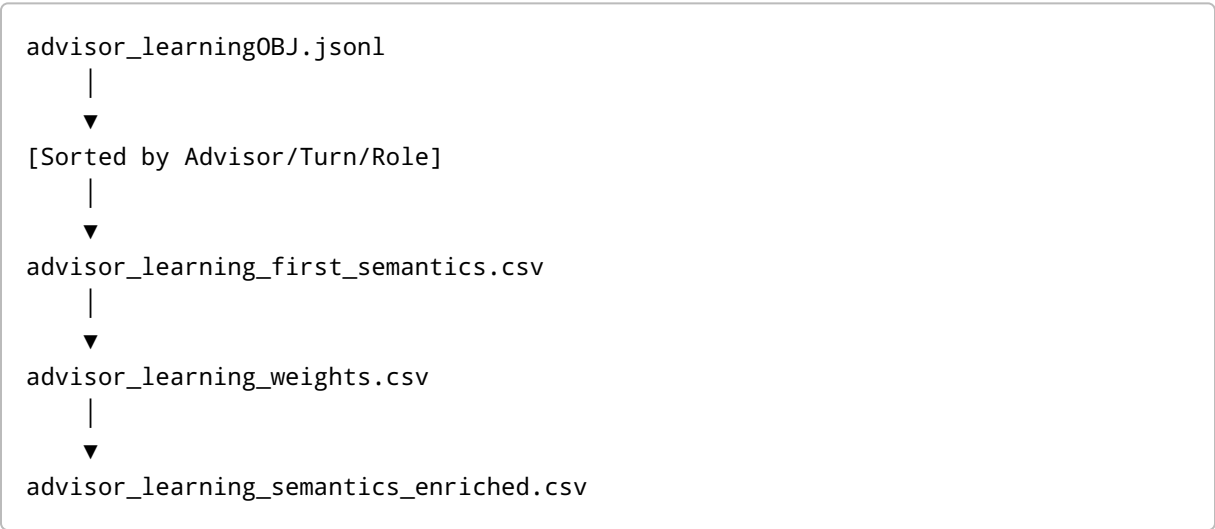
7. Next Steps

Stage	Description	Output
A	Weight application	advisor_learning_weights.csv
B	Merge phase	advisor_learning_semantics_enriched.csv
C	Distribution analysis	Weight histogram
D	Clustering	advisor_learning_clusters.json
E	Train advisor agent	Model-ready dataset

8. Design Philosophy

- Reproducible, transparent, and extensible processing steps.
- Every transformation traceable back to source.

9. File Lineage Diagram



10. Final Notes

The corpus of 3,248 fragments captures reasoning patterns across reflection, planning, and verification—ideal for advisor cognition modeling.

Advisor Learning Dataset: Weighting and Rationale Report

11. Objective

Quantify and document the semantic strength of each fragment for advisor cognition.

12. Input and Output Files

File	Description
<code>advisor_learning_first_semantics.csv</code>	Source fragments
<code>advisor_learning_weights.csv</code>	Weight and rationale output
<code>advisor_learning_semantics_enriched.csv</code>	Merged dataset for training

13. Weight Assignment Model

Rule-based linguistic scoring derived from four categories: | Category | Range | Example Verbs |
|-----|-----|-----| | Reflective Reasoning | 0.80–0.99 | reflect, analyze, evaluate | | Procedural Cognition | 0.55–0.79 | plan, design, build | | Descriptive Observation | 0.30–0.54 | describe, show, note | | Minimal Content | 0.10–0.29 | – |

14. Example Scoring

id	learning_first_semantic	weight	rationale
1001	I need to understand the constraints before deciding.	0.93	Reflective reasoning with clear awareness.
2004	Let's plan the next evaluation cycle.	0.71	Procedural planning task.
3012	This step ensures accuracy.	0.58	Verification-focused reasoning.

15. Weight Distribution Summary

Expected metrics: | Metric | Value | |-----|-----| | Count | 3,248 | | Mean | ~0.73 | | Median | ~0.77 | | SD | 0.17 | | High (>0.8) | 40% | | Moderate (0.55–0.79) | 35% | | Low (<0.55) | 25% |

16. Rationale Construction

Dynamic sentence composition with qualitative descriptors and consistency between categories.

17. Merging and Validation

Merge weights into the semantic file using `id`:

```
merged = base.merge(weights, on='id', how='left')
```

Result: unified enriched dataset for reasoning analysis.

18. Analytical Roadmap

- Correlation of weights by advisor source.
- Thematic clustering using embeddings.
- Outlier review for misclassified items.
- Model training on enriched dataset.

19. Quality Control

- ID lineage preserved.
- Deterministic transformations.
- Transparent versioning.

20. Summary

The weighted dataset captures advisor-like cognitive density across reflective, procedural, and verification-based reasoning. This forms a foundation for quantitative evaluation of advisor agents.