

Supported or Not: The System One Dataset Specification

Version 1.0*

Logitly AI
research@logitly.ai

Abstract

This document specifies System One, a dataset that trains and evaluates decision models on a single judgment: is a candidate answer supported by the given state? Such models make fast, structured judgments that software consumes directly: a choice or a probability, never prose. Truth is relative to the state, not the world: a candidate is supported when the state entails it, regardless of real-world fact. We define the record format and label semantics, a verification protocol, the supported task constructions, and three baselines a model must beat before its score means anything.

Keywords

System One, Decision Making, Verified Dataset

1 Overview

A growing class of decision models serves AI agents rather than human readers: given a context and a question, they return a typed choice or a calibrated probability, with no generated text to parse. System One trains and evaluates such models on one judgment: *is a candidate answer supported by the given state?* A model reads a state, a query, and a candidate; it emits a choice or a probability. It never generates text. The role is the fast, intuitive counterpart to a generative model’s slow, deliberate one, and it is useful only if the fast answer can be trusted to follow the provided context.

That trust is what this dataset provides. The specification rests on one invariant: **truth is relative to the state, not the world**. A candidate is *supported* when the state entails it, regardless of whether it is true outside the record. This rule separates grounding from fact-checking, makes every record decidable from the state alone, and fixes what every label asserts: faithfulness to provided context, nothing else.

Records are JSON Lines: UTF-8, one record per line, in English. Each record has five fields: an identifier, a state, a query, and two candidates, *positive* and *negative*. The state describes the facts that the query asks about. *positive* and *negative* are two candidate answers: one is supported by the state, one is not. Table 1 shows one record.

2 Label Semantics

2.1 Grounding

A candidate is supported when every claim it makes is entailed by the state. Three consequences:

- **Truth is relative.** The label describes the candidate’s relation to the state, never its relation to the world.
- **Rewording is allowed.** Meaning-preserving paraphrase does not break support.

Table 1: A record from the dataset.

id	jev-3414003499
state	Free shipping is available for all U.S. orders over \$49. Regular shipping costs \$5.95. Deliveries take 3 to 5 business days in the contiguous United States. Deliveries to Alaska, Hawaii, and P.O. boxes may take longer. Deliveries to APO/FPO addresses may take up to 21 days. Orders placed after 12pm EST will be shipped the next business day.
query	What are the shipping costs and delivery times for U.S. orders?
positive	Free shipping is available for all U.S. orders over \$49, while regular shipping costs \$5.95. Deliveries take 3 to 5 business days in the contiguous United States, though Alaska, Hawaii, and P.O. boxes may take longer, and APO/FPO addresses may take up to 21 days.
negative	Free shipping is available for all international orders over \$49, while regular shipping costs \$5.95. Deliveries take 3 to 5 business days in foreign countries, though remote areas may take longer, and military addresses may take up to 21 days.

- **Addition is not allowed.** An added causal link or outside detail breaks support.

2.2 Positives

A positive (i) is supported under the rule of §2.1 and (ii) answers the query. It may be partial: it need not cover everything the query asks (Table 1). Completeness is deliberately out of scope (§7).

2.3 Negatives

A negative is topically related to the query and the state, and is usually written in a form similar to the positive; it is the false answer. Each negative is exactly one of two kinds:

- **Contradiction.** It conflicts with at least one claim in the state: a changed number, entity, polarity, or scope, or a swapped direction or role.
- **Weak neutral.** It adds claims the state neither supports nor refutes.

The kind is not labeled, by design. The dataset targets one binary judgment: supported or not. Exposing the kind would invite models to solve an easier, NLI-shaped task instead of checking the state.

2.4 Difficulty

Negatives range from easy (wholesale rewrites, several obvious conflicts) to hard (a single minimal edit). Difficulty is not labeled, by design.

*This document is a dataset specification, not a research paper; do not distribute.

Table 2: Task construction.

Task	Construction	Target
Binary	state + query + one candidate	true or false
Choice	state + query + both candidates, order shuffled	the positive
Rerank	state + query + both candidates, triplet	positive ranked above negative

3 Verification Protocol

Every record is machine-verified. A record is valid only if all checks hold:

- **Positive is supported.** Every claim in the positive is entailed by the state, with no added causation and no outside knowledge.
- **Negative is not supported.** The negative contradicts the state, or adds claims the state does not support.
- **Factual consistency.** Every number, date, and named entity in the positive matches the state.
- **Distinctness.** Positive and negative differ in *meaning*, not merely in wording. A negative that is a near-copy of, or a subset of, the positive is invalid.
- **Relevance.** The positive answers the query; the negative is topically related to it.
- **Self-containment.** The state alone suffices to judge both candidates.
- **Uniqueness.** No duplicate records, and no overlap with public benchmarks.

4 Intended Usage

Table 2 lists the three supported constructions. Each construction puts one typed question to a decision model, with the record’s state as the model’s state input.

- **Binary.** It judges one candidate, *is this supported by the state?* The answer is a confidence indicating positive or negative.
- **Choice.** It scores both candidates: positive must score higher than negative, whichever order they appear in.
- **Rerank.** It is the one construction that is not a question: training and evaluation data for rankers and embedders, which must place positive above negative.

5 Known Characteristics

How negatives are built. Observed construction patterns, unlabeled: number or date change; entity or subject swap; polarity flip; scope or quantifier change (e.g., *only*, *universally*); wholesale opposite claim; shared prefix with one edited part; off-target unsupported claim.

Training-time risks.

- **Cue words.** Some negatives carry absolute or contrastive markers (*cannot*, *only*, *lacks*, *excludes*, *did not*, *universally*). A model can learn the marker instead of checking the state.
- **High lexical overlap.** Both candidates typically reuse the state’s wording. On easy records, overlap alone rejects the

negative; on hard negatives the two candidates’ overlap is nearly identical, so overlap is not a reliable signal anywhere.

- **Easy negatives.** Some negatives can be rejected without reading the state at all.

Reading the data.

- **Partial positives.** A positive may omit part of what the query asks. In Table 1, the positive omits the next-business-day shipping cutoff; another record omits the \$5.95 regular shipping cost and the longer Alaska/Hawaii/APO timelines. The dataset measures faithfulness, not completeness.
- **Query style.** Queries follow similar templates and often restate the state closely; they do not span the variety of real user queries.
- **Edit location.** The distinguishing edit can sit anywhere in a candidate, including the final sentence, which is why evaluation must not truncate candidates (§6).

6 Recommended Evaluation

- **Metrics.** One per construction (Table 2). For Binary, AUROC on the probability the model reports. For Choice, accuracy of picking the positive; for Rerank, the fraction of records with the positive ranked above the negative.
- **Baselines.** Report three: length, lexical overlap, and score without state. A model’s score is meaningful only if it clearly beats all three; otherwise it is exploiting length, overlap, or query phrasing rather than the state.
- **Order robustness.** In the Choice construction, swapping the two candidates must not change which one the model picks; report the delta.
- **Adjustable threshold.** A model trained on this dataset should make its threshold adjustable. One way is to apply a temperature to the raw score before it becomes the confidence; the rescaling is a monotonic transformation, so AUROC is unaffected.
- **No truncation.** Do not truncate candidates: the only distinguishing edit may fall in the final sentence, which truncation removes first.
- **Interpretation.** High accuracy alone does not establish grounding. Read it against the baselines.

7 Out of Scope

The dataset does not measure answer completeness, does not fact-check against world knowledge, and does not evaluate free-text generation. All three exclusions follow from the invariant of §1: judgment is relative to the state and reported as a choice or a probability.

Complete characterization is out of scope: the specification is a post-hoc, human-written account of AI-generated data. Records may exhibit characteristics that are unknown or undocumented here.