



Generated — OSHA_29CFR_full_2024 — veritrooper v9
13 July 2026

Datasheet for the purpose-generated evaluation set used in this audit. Generated 2026-07-13 23:17:27.

Motivation

This question set exists to measure the model's accuracy and robustness on the provider's own material — not on a generic public benchmark. Questions are generated from the source document so the evaluation reflects the deployment domain.

Composition

Field	Value
Total items	1000
Categories	7
Refusal / unanswerable (hallucination-trap) items	200
Source material	data/OSHA_29CFR_full_2024.txt

Items per category:

Question type	Items
Cross-Reference	251
Precision	218
Conditional	211
Negative (hallucination traps)	200
Exception	93
Cause & Effect	15
Calculation	12

Generation process

- The source document was chunked; questions were generated per chunk with code-computed ground truth where the answer is arithmetic or tabular.
- Adversarial negative items (hallucination traps) were generated to have **no supported answer** — the correct response is to refuse — and passed through a verification gate before inclusion.
- Items whose ground truth was later found malformed are quarantined as **bad tests** and excluded symmetrically from scoring rather than counted against either arm.

Recommended uses & limitations

- **Use:** point-in-time accuracy/robustness evaluation of a model on this material; regression testing across model or prompt changes (re-run the frozen panel).
- **Not for training.** These items are evaluation artifacts; training on them would contaminate future measurement.
- **Not a traffic sample.** The set probes capability by question type; it is not a statistical sample of real production queries, and category counts are generation-driven, not usage-weighted.

Provenance & reproducibility

Field	Value
Engine version	VERITROOPER v9
Generated data file	data/OSHA_29CFR_full_2024_generated_20260601_104738.json
Generation temperature	0.1
Sampling strategy	full

Contents (click a section to jump)

Motivation

Composition

Generation process

Recommended uses & limitations

Provenance & reproducibility

Generated by VERITROOPER as a data card documenting the evaluation test set. This artifact is testing evidence only. It does not by itself constitute, certify, or confer conformity with any framework, and does not replace the conformity assessment or the provider's other obligations. Have counsel review before external use.

