

Halvard Steel AG 2025 — EUR 2058 Notes

CDSBench #20 · Does the record contain enough evidence to reach a reliable deliverability conclusion without further documents?

Halvard Steel AG · Corporate · Restructuring 2025-02-04 · Uncertainty · Uncertainty · Ground truth: deterministically verified

VERIFIED OUTCOME

Yes

A deliverable-obligation characteristic fails on the record.

The label follows from the whole record.

WHAT HAPPENED

All 4 participants reached the verified outcome.

Decided by analyst: The relevant fact is not stated in the provided documents.

MODEL COMPARISON

PARTICIPANT	ANSWER	RESULT	CONFIDENCE	EVIDENCE SCORE	PRIMARY FAILURE
Claude	Yes	Correct	40%	100%	—
GPT	Yes	Correct	92%	100%	—
Gemini	Yes	Correct	85%	100%	—
Mastermind (reference)	Yes	Correct	40%	100%	—

EVIDENCE

Decisive sections of the record

METHODOLOGY

A response is correct when its normalised answer equals the verified label. Evidence score is the F1 between cited section IDs and the supporting sections; cited IDs that do not exist count as hallucinated. Abstentions are correct only when the label calls for them. Ground truth is verified by two independent extractors plus an evidence check; no model grades itself.

Full Case Replay with documents and every model's reasoning: <https://cdsbench.com/cases/halvard-steel-ag-2025/eur-2058/sufficiency>