

AI AGENT EVALUATION

Online store support agent

A customer is still waiting for a refund. The record shows approval, but no completed payment.

Completed sample5 human reviewers3 criteria · 1-5

Overall resultNeeds attention

2.7/5

Mean human grade for the original reply.

Key findings

AI assessment

- Your refund is approved.

Refund record · status: approved

Supported
- The payment is on its way.

Payment confirmation · not available

Unverified
- You're all set.

Customer request · arrival date still unanswered

Unsupported

Evaluation details

- ReportDEMO-01 · illustrative
- ScenarioRefund follow-up
- Agent categoryShopping & retail
- Review typeAI assessment + human review
- Human reviewers5 fictional reviewer records
- ResponseA · original reply

Category grades

Human review

- Correctness2.2/5
- Helpfulness2.0/5
- Tone3.8/5

Reviewer grades

Original reply · grade / 5

Individual judgments from 5 illustrative reviewers.

Reviewer	Correctness	Helpfulness	Tone	Mean	Prefers
Reviewer 01	2	2	4	2.7	B
Reviewer 02	2	2	4	2.7	B
Reviewer 03	3	3	4	3.3	A
Reviewer 04	2	2	3	2.3	B
Reviewer 05	2	1	4	2.3	B
Average	2.2	2.0	3.8	2.7	80% B

Sources & evidence

- 1Refund record · status: approved
Illustrative evidence · E1
- 2Payment confirmation · not available
Illustrative evidence · E2
- 3Customer request · arrival date still unanswered
Illustrative evidence · E3

Report contents

- A. Conversation & evidencePage 2
- B. Human review & preferencesPage 3
- C. Criteria & interpretationPage 4

Illustrative report. No actual company or agent performance is represented.

Conversation and evidence

A customer is still waiting for a refund. The record shows approval, but no completed payment.

Customer conversation excerpts

6 turns

Customer · turn 1

"Could you check the status of my refund?"

Customer · turn 2

"Thanks. When should I expect it?"

Customer · turn 3

"I was told it was approved last week."

Customer · turn 4

"That doesn't tell me whether the money was sent."

Customer · turn 5

"Can someone check the actual payment?"

Customer · turn 6

"So I still don't know when it will arrive?"

Customer excerpts are shown here; these are not a complete turn-by-turn transcript.

RESPONSE A · UNDER REVIEW

Your refund is approved. The payment is on its way. You're all set.

AI sentiment estimate

-38 points

1

2

3

4

5

6

Estimated customer sentiment: 62 → 24.

Scale: 0 = frustrated / 100 = positive. This is an AI estimate, separate from human grading.

Supporting evidence

3 claim checks

E1

"Your refund is approved."

The refund record shows an approved request. This part of the answer matches the available evidence.

Refund record · status: approved

Supported

E2

"The payment is on its way."

The available record contains no payment confirmation. Approval alone does not establish that a transfer has started.

Payment confirmation · not available

Unverified

E3

"You're all set."

The customer asked when the money would arrive. The reply presents the issue as resolved without an arrival date or a confirmed payment.

Customer request · arrival date still unanswered

Unsupported

Overlap.

Illustrative data · fictional conversation, evidence and human reviews

2 / 4

Human review and reply preferences

Grades assess the original reply. Preference is a separate choice between replies A and B.

80%

4 of 5 reviewers chose the alternative reply. Individual assessments, including dissenting judgments, are listed below.

preferred reply B

Reviewer 01

Correctness 2/5 Helpfulness 2/5 Tone 4/5

Prefers B

"B makes the uncertainty clear and offers a next step. A sounds finished when the customer's problem isn't."

Reviewer 02

Correctness 2/5 Helpfulness 2/5 Tone 4/5

Prefers B

"I would rather be told what is known and what still needs checking than be reassured without evidence."

Reviewer 03

Correctness 3/5 Helpfulness 3/5 Tone 4/5

Prefers A

"I prefer A's brevity and directness, although it should include a realistic payment timeframe."

Reviewer 04

Correctness 2/5 Helpfulness 2/5 Tone 3/5

Prefers B

"B separates approval from payment and gives the customer a practical route forward."

Reviewer 05

Correctness 2/5 Helpfulness 1/5 Tone 4/5

Prefers B

"A repeats a status the customer already knows. B acknowledges the unresolved part of the request."

A · Original reply

1 reviewer

Your refund is approved. The payment is on its way.
You're all set.

B · Alternative reply

4 reviewers

Your refund is approved, but I can't confirm it has been sent yet. I can ask the payments team to check and give you an update.

Assessment criteria and interpretation

The criteria below are illustrative. Evaluation criteria and supporting context are specified for each task.

Correctness
Does the response match the supplied context and evidence?

1 / Low Material claims are contradicted or unsupported.	3 / Mixed Partly grounded, with an important omission or unresolved uncertainty.	5 / Strong Claims match the evidence and uncertainty is made clear.
---	---	--

Helpfulness
Does the response address the customer’s request and help them move forward?

1 / Low The request remains unresolved and no useful next step is offered.	3 / Mixed Part of the request is addressed, but the next step needs more detail.	5 / Strong The customer receives a relevant answer and a practical way forward.
---	---	--

Tone
Does the response communicate clearly and acknowledge the customer’s situation?

1 / Low Dismissive, confusing or inappropriately reassuring.	3 / Mixed Clear and civil, but generic or insufficiently attentive.	5 / Strong Respectful, attentive and appropriately confident.
---	--	--

Result interpretation

Human grades	Each category average uses all 5 reviewer grades. The overall mean gives the three categories equal weight and is displayed to one decimal place.
Reply preference	The preference percentage counts the chosen reply. It does not establish that the preferred reply is factually correct.
AI claim checks	Supported means the supplied evidence backs the claim. Unverified means confirmation is missing. Unsupported means the claim is contradicted or goes beyond the available evidence.
Sentiment drift	The change between the first and final AI-estimated customer sentiment values. It is distinct from human grades and reply preferences.

Sample data disclosure
All conversations, evidence records, model assessments, grades and reviewer judgments in this document are fictional examples. They demonstrate the report format and are not evidence of Overlap’s track record or a real company’s agent performance.