

Bilingual AI Response Evaluation

Case 04 Portfolio simulation

Xuanhui Zhang | Budapest | sallaizhang@outlook.com

Prepared 8 October 2026

This simulation shows how I review a response against a reference, score its quality and decide whether to accept, revise, reject or escalate. Six English and Chinese cases include factual errors, incomplete answers, localization choices and unresolved task ambiguity.

Evaluation workflow

Read the task and fact card; check factual statements and numbers; score each dimension; apply the critical-error override; record a preferred candidate and action; and explain the correction or clarification needed. Preferences do not automatically imply acceptance.

Rubric

Dimension	Weight	Rating focus
Grounding G	40%	Accuracy against the fact card; numbers and meaning.
Instruction following I	25%	Completion, requested format and constraints.
Language and localization L	20%	Clarity, grammar, terminology and audience.
Calibration C	15%	Scope, uncertainty and appropriate inference.

Score each dimension 0–4: 4 fully meets the criterion; 3 minor defect; 2 material but repairable omission; 1 major defect; 0 contradiction or task failure. Total = $25 \times (0.40G + 0.25I + 0.20L + 0.15C)$.

Accept: at least 85 and no critical error. Revise: 60–84.9 and no critical error. Reject: below 60 or any critical error. A fabricated fact or material inversion of a number, missing value or eligibility status is critical. Escalate when source ambiguity prevents a justified answer; a provisional score cannot resolve missing instructions. Ties favor stronger grounding.

Client deliverables

A rubric, compact score sheet, case-level rationale, error tags, corrections and an unresolved-case queue. The worked examples below show both clear choices and cases where the task needs more information.

English evaluation cases

E01

Fact card: Hungary 2024 Germany export share = 0.2463004459595256 of eligible goods exports. Task: one sentence with the percentage and a descriptive limitation.

Candidate A: Germany accounted for 24.6% of Hungary's eligible goods exports in 2024; this measures trade exposure, not political dependence.

Candidate B: Germany controlled 24.6% of Hungary's economy in 2024, proving political dependence.

G I L C — A (4, 4, 4, 4): 100.00; B (0, 1, 4, 0): 26.25. Preferred: A. Decision: Accept. Tags: NUM CAU.

A preserves the denominator, year and limitation. B changes goods exports into the whole economy and infers political control; reject B as critical.

E02

Fact card: Slovakia 2024 import eligible coverage = 0.891495; threshold = 0.95. Task: include or exclude import shares, then state the next action.

Candidate A: Exclude the import shares: 89.1495% eligible coverage is below the 95% threshold. Keep the failed group in the log and review the coverage gap before comparing imports. This should be reviewed before further comparative analysis is conducted.

Candidate B: Include the group because raw coverage is enough; missing partner values are zero.

G I L C — A (4, 4, 3, 4): 95.00; B (0, 0, 4, 0): 20.00. Preferred: A. Decision: Accept. Tags: LOC GATE MISS.

A earns 95: the decision and next action are correct, with one redundant sentence. B bypasses the gate and assigns zero without evidence; reject B as critical. Edit A for concision.

E03

Fact card: 980 observation records; 67 null-valued; 65 pending; categories may overlap. Task: summarize the total, both counts and the overlap caution in no more than two sentences.

Candidate A: The export contains 980 records, including 67 null-valued records and 65 pending records. These categories may overlap and should not be added together.

Candidate B: The export has 980 records, including null-valued and pending records. Both categories should be checked.

G I L C — A (4, 4, 4, 4): 100.00; B (3, 2, 4, 2): 70.00. Preferred: A. Decision: Accept. Tags: SCOPE.

A earns 100: preserves the total and both counts, states the overlap caution explicitly and avoids false addition in two concise sentences. B earns 70: it omits both counts and the overlap condition. Accept A; B requires revision.

Chinese and localization evaluation cases

Z01

事实卡：德国出口份额为 24.6%，仅描述贸易敞口。任务：用一句自然的中文写出份额，并明确说明不能据此判断政治依赖。

Candidate A: 德国占出口额的 24.6%，这只是贸易敞口的一个描述。不能据此判断政治依赖。

Candidate B: 德国占出口额的 24.6%。

G I L C — A (4, 3, 3, 4): 88.75; B (4, 2, 4, 2): 80.00. Preferred: A. Decision: Accept. Tags: INST LOC.

A earns 88.75: preserves the fact and limitation, but exceeds the one-sentence instruction and uses slightly awkward Chinese. B omits the limitation. Correct to “德国占出口额的 24.6%，这一份额描述贸易敞口，不能据此判断政治依赖。”

Z02

事实卡：缺失值为 null，原因是尚未公布。任务：写明当前处理、缺失原因和何时更新，且不填零。

Candidate A: 保留缺失值，待正式数据公布后再更新。

Candidate B: 保留缺失值，不填零。

G I L C — A (3, 2, 4, 3): 73.75; B (3, 2, 4, 2): 70.00. Preferred: A. Decision: Revise. Tags: MISS SCOPE.

Neither response fully satisfies the task. A earns 73.75 and omits the explicit “not yet published” reason and zero policy. B omits the reason and update trigger. Prefer A for its update instruction; revise to “保留缺失值及‘尚未公布’的原因，不填零；待正式数据公布后再更新。”

Z03

事实卡：界面源文为“提交后 15 分钟内处理”。没有定义“处理”是开始审核还是完成审核。任务：给出可发布英文，明确处理完成的承诺，并评估是否需要追问。

Candidate A: Your submission will be fully processed within 15 minutes.

Candidate B: Your submission will be processed within 15 minutes; please confirm whether this means review starts or finishes.

G I L C — A (1, 2, 4, 0): 42.50; B (3, 2, 4, 4): 77.50. Preferred: Neither. Decision: Escalate. Tags: AMB SCOPE.

A earns 42.50 and invents a completion commitment; reject A as critical. B earns 77.50 and correctly raises the ambiguity, but its first clause still risks implying completion and is not publication-ready. Ask the owner what “处理” guarantees, then draft either “Review will begin...” or “Review will be completed...”. Neither response fully satisfies the task.

Score sheet error taxonomy and handover

Case	A score	B score	Preferred	Decision	Main tag
E01	100.00	26.25	A	Accept	NUM
E02	95.00	20.00	A	Accept	LOC
E03	100.00	70.00	A	Accept	SCOPE
Z01	88.75	80.00	A	Accept	INST
Z02	73.75	70.00	A	Revise	MISS
Z03	42.50	77.50	Neither	Escalate	AMB

Scores are out of 100. Decision applies to the preferred candidate, or to the pair when escalated. Critical rejected candidates are identified in each case rationale. E03 meets all four criteria after revision; Z03 remains escalated regardless of provisional scoring.

Tag	Error type	Review response
NUM	Number or denominator error	Check against the exact fact-card quantity.
CAU	Unsupported causal claim	Remove inference unsupported by the reference.
GATE	Eligibility inversion	Check threshold and group decision.
MISS	Missing-value handling	Preserve missingness and reason.
SCOPE	Lost limitation or missing context	Restore scope or request clarification.
LOC	Localization or language defect	Correct wording and terminology.
INST	Instruction or format defect	Apply the stated task constraints.
AMB	Ambiguous source instruction	Escalate before approving publication.

Quality control and what this demonstrates

The score sheet is reconciled to the weighted formula and case rationales. Acceptance with minor defects, revision and escalation stay distinct. In a client workflow, I can review bilingual responses against supplied references, document evidence-linked preferences and maintain an issue queue. Excel can support row-level scores and review logs.

Scope and evidence note

All prompts, candidate answers and scores are simulated portfolio exercises. E01–E03 use pinned Atlas facts; Z01–Z03 use authored fact cards. The pack demonstrates review decisions rather than model benchmarking or inter-rater reliability. Atlas snapshot 994d61fe1e56, reviewed 8 October 2026.

[src/data/network/network_metrics.json](#)

[src/data/network/network_coverage.json](#)

[public/research-data/observations.json](#)