LANGUAGE TECHNOLOGY / QUALITY ASSURANCE

AI-Assisted Language Content Quality Assurance

A practical, evidence-conscious workflow for designing, validating and reviewing English listening lessons for German-speaking adult learners.

3 lesson packages | 12 test specifications | 4 executed assertions

Independent portfolio study · A2-targeted listening content · English / German

Not affiliated with Remba or any commercial language-learning product. No official CEFR certification or learner-outcome claims.

CASE STUDY / 01

Executive summary

The project demonstrates a small, reproducible content production workflow rather than a finished language-learning platform.

The problem

Generative AI can draft plausible educational dialogues quickly, but fluency alone does not establish grammatical accuracy, suitable difficulty, answerability, instructional value or audio readiness. A repeatable QA process is needed before materials are released to learners.

The proposed response

• Write a consistent lesson specification with bilingual support instructions and a traceable question-to-dialogue link.

• Check measurable format rules automatically; reserve semantic and pedagogical decisions for qualified human review.

• Track defects, revisions and release decisions. Document where evidence is missing rather than assuming the system worked.

What is actually demonstrated

Deliverable	Evidence status
Three packaged dialogue lessons	All three pass the Python deterministic validator.
12 test specifications	Four automated assertions executed; eight require human review.
Prompt revisions v1.0 and v1.1	Documented variants only; no controlled comparison.
Learner or audio outcomes	Not tested; no audio files or learner study.

CASE STUDY / 02

Audience, learning goals & scope

A constrained content format keeps the learning experience and the QA criteria observable.

Learner profile

Adult German-speaking learners of English whose learning goals approximately align with selected A2 listening situations. The German language is used for short instructions and feedback; English is used for the dialogue and comprehension items.

Core task

After listening to a short everyday exchange, learners identify explicit facts such as a time, a request, or a workplace activity. The design prioritizes practical interaction and accessible listening input.

Included vs excluded

In scope	Out of scope
Three dialogue-based lessons; structured questions and answers; glossary entries.	Mobile application, speech recognition, interactive audio player.
Prompt design, JSON model, deterministic validator, test fixtures.	External production system, model benchmark, comprehensive CEFR mapping.
Editorial criteria, issue tracking and human review process.	Independent review sign-off, learner efficacy or audio intelligibility study.

Definition of done

A lesson is structurally valid, its correct answers are independently confirmed from the dialogue, the language is appropriate to the intended scenario, and an authorized reviewer records approval. This portfolio currently meets the structural part only; its packaged “approved” status is illustrative.

CASE STUDY / 03

Content production workflow

The pipeline separates creation, rules-based checks and judgement-dependent editorial review.

01 · Lesson specification

Set topic, learner needs, learning objectives and output constraints.

02 · Draft content

Use versioned prompts to create structured lesson JSON.

03 · Deterministic validation

Reject invalid format, missing data or measurable constraint violations.

04 · Linguistic review

Verify accuracy, naturalness, level suitability, translations and pedagogy.

05 · Revision and regression

Document edits, rerun relevant tests and assess new errors.

06 · Release decision

Record reviewer identity, approval rationale and remaining risk.

Automated checks do not substitute for bilingual proofreading, nor can a reviewer infer real audio clarity from an unrecorded script.

CASE STUDY / 04

Lesson architecture

Each sample lesson follows the same predictable machine-readable shape.

Illustrative JSON fields

Field	Function
lesson_id / version	Traceability across revisions.
target_language / support_language / level	Language pair and target level label.
learning_objectives	Expected communicative outcomes.
audio_instructions_de	Concise German listening-task instructions.
dialogue	Ordered speaker-labelled utterances.
questions	Three comprehension items with answer, evidence_quote and German feedback.
glossary / status / review	Vocabulary support and editorial metadata.

Constraints used in the project

- 8–12 speaker turns and 90–130 spoken dialogue words, based on internal design choices.
- Three answerable comprehension questions with evidence quotations.
- Explicit speaker labels, consistent entities and clear numeric references.
- The scripts should be read aloud naturally and independently of visual cues.

These numeric constraints are project conventions; they are not official CEFR requirements.

CASE STUDY / 05

Lesson 01: Ordering at a Café

A concrete illustration of the content, task design and traceability built into the structured lesson package.

Learning objectives

- Recognise polite ordering phrases
- Identify an order and a price from a short exchange

Dialogue extract

Barista: Good morning! Welcome to Green Street Café. What can I get for you today?

Customer: Good morning. Could I have a small coffee with milk, please?

Barista: Of course. Would you like it hot or iced?

Customer: Hot, please. And do you have any cheese sandwiches?

Barista: Yes, we do. They come with tomatoes and fresh bread.

Customer: That sounds good. I would like one sandwich as well.

Question-to-evidence mapping

Comprehension question	Expected answer	Dialogue evidence
What drink does the customer order?	A small hot coffee with milk.	Could I have a small coffee with milk, please?
How much is the whole order?	Eight euros.	So that is eight euros altogether, right?
Does the customer eat at the café?	Yes, the customer eats there.	I will eat here, thank you.

Shown is an excerpt for legibility. The full lesson dialogue, glossary, German feedback and all metadata are included in the ZIP project files.

CASE STUDY / 06

Lesson 02: Planning a Train Journey

A concrete illustration of the content, task design and traceability built into the structured lesson package.

Learning objectives

- Identify a departure time and platform
- Ask for and understand ticket details

Dialogue extract

Passenger: Hello. I need to travel to Brighton on Saturday morning.

Clerk: Certainly. Are you travelling from London Victoria station?

Passenger: Yes. I would like to arrive before eleven o’clock.

Clerk: There is a direct train at nine fifteen. It arrives at ten twenty.

Passenger: That sounds perfect. How much is a return ticket?

Clerk: A return ticket is twenty-four pounds for one adult.

Question-to-evidence mapping

Comprehension question	Expected answer	Dialogue evidence
When does the direct train leave?	At 9:15.	There is a direct train at nine fifteen.
How much is one adult return ticket?	Twenty-four pounds.	A return ticket is twenty-four pounds for one adult.
Which platform does the train usually leave from?	Platform six.	It usually leaves from platform six, but please check the screen.

Shown is an excerpt for legibility. The full lesson dialogue, glossary, German feedback and all metadata are included in the ZIP project files.

CASE STUDY / 07

Lesson 03: Talking About Work

A concrete illustration of the content, task design and traceability built into the structured lesson package.

Learning objectives

- Describe everyday work tasks
- Identify a meeting time and short-term work plans

Dialogue extract

Mia: Hi, Daniel. How is your new job at the bookshop going?

Daniel: It is going well, thanks. I start at nine every morning.

Mia: What do you usually do when you arrive?

Daniel: First, I put new books on the shelves and check online orders.

Mia: Do you also help customers find books?

Daniel: Yes. That is my favourite part of the job.

Question-to-evidence mapping

Comprehension question	Expected answer	Dialogue evidence
What time does Daniel start work?	At nine in the morning.	I start at nine every morning.
When is the team meeting?	Thursday at two.	We have a team meeting on Thursday at two.
What event will the team plan?	A small reading event for local families.	We will plan a small reading event for local families.

Shown is an excerpt for legibility. The full lesson dialogue, glossary, German feedback and all metadata are included in the ZIP project files.

CASE STUDY / 08

Quality rubric & release gates

A weighted review rubric makes judgement transparent but does not remove the need for qualified editorial evidence.

Criterion	Weight	Reviewer focus
Linguistic accuracy	25	Grammar, spelling, semantics and lexical usage.
Level appropriateness	20	Accessibility for selected A2-aligned goals.
Pedagogical value	20	Purposeful tasks, answerability and useful feedback.
Consistency & naturalness	15	Realistic speech and stable facts.
Instruction compliance	10	Schema, counts and explicit constraints.
Audio readiness	10	Oral clarity, pronunciation risks and audible distinctions.

Critical blocking issues

- Incorrect language modelled as correct or a materially misleading translation.
- A comprehension answer not supported by the dialogue.
- Mismatch between recorded audio and answer key (if audio exists).
- Invalid JSON or missing fields that prevent ingestion.

A numerical score is an editorial aid, not sufficient grounds for release where a critical issue is present. No independently completed rubric scores are claimed in this package.

CASE STUDY / 09

Test strategy & controlled defects

The 12 fixtures cover both conditions that code can verify and defects that need linguistic or pedagogical interpretation.

ID	Risk / expected control	Execution
TC-01–02	Verb tense and word order	Manual
TC-03–04	Politeness / level mismatch	Manual
TC-05–06	Translation error / ambiguous item	Manual
TC-07	Evidence quotation absent	Automated pass
TC-08	Inconsistent time or price	Manual
TC-09	Invalid JSON	Automated pass
TC-10	Missing topic field	Automated pass
TC-11	Audio-unfriendly script	Manual
TC-12	Valid clean control	Automated pass

Interpretation boundary

TC-07 verifies whether a stated quotation occurs in the script. It does not prove that an answer is semantically entailed by the text. Likewise, the linguistic fixtures are synthetic examples, not measured mistakes from an external model run.

Observed results

Evidence is intentionally reported as a small technical demonstration, with measured results separated from planned evaluation.

Metric	Observed result	Interpretation
Packaged lesson validation	3 / 3 pass	Structural validation only.
Defined test cases	12	Coverage design, not 12 automated checks.
Executed assertions	4 / 4 pass	Narrow deterministic behavior confirmed.
Human-only cases	8 pending	No independent reviewer judgement logged.
v1.0 vs v1.1 performance	Not measured	Cannot infer quality gains.
Review time / learner impact	Not measured	No speed or educational efficacy claims.

What these results establish

For the supplied example fixtures, the deterministic tests detect selected structural problems and accept a structurally valid control. Three packaged lessons satisfy the validator’s implemented constraints. This is evidence of a functioning validation script, not of end-to-end linguistic reliability.

What the results do not establish

No precision/recall, model error rate, human time saving, CEFR attainment, multilingual generalizability or production readiness can be inferred from this evaluation.

CASE STUDY / 11

Human review in practice

Editorial decision-making should be explainable and auditable, especially when a rule is linguistic rather than purely structural.

A controlled editorial example

Original — deliberately injected	Recommended edit
“I want that you give me a coffee.”	“Could I have a coffee, please?”

The original is an intentionally constructed interference example. It is not represented as a real model output. The revision replaces a non-idiomatic request with a naturally polite phrase appropriate for a café scenario.

Editorial reviewer checklist

- Distinguish grammatical mistakes from stylistic preference and contextual appropriateness.
- Verify every comprehension answer against the full dialogue, not only the evidence_quote string.
- Check German support feedback for meaning preservation.
- Read the script aloud or assess the synthesized recording; document numbers, names and speaker differentiation.
- Log the issue severity, exact text span, proposed revision and regression test.

Prompt governance & next iteration

Prompt quality must be evaluated through comparable model runs, not through documentation alone.

Two design variants

Generation v1.0	Generation v1.1
Basic constraints on scenario, difficulty, format and question count.	Stronger instruction to link each correct answer to explicit source evidence and reject unsupported inference.

Both versions are supplied as prompt documents, but there are no paired API generation runs under fixed parameters. Their relative performance is unknown.

Proposed reproducible experiment

- Generate at least 30 unseen dialogue outputs per controlled prompt condition with recorded model and inference settings.
- Obtain independent bilingual annotations and adjudicate disagreements before scoring.
- Track per-category misses, false flags, release blockers and human-review time.
- Test actual audio from a specified TTS configuration and collect consent-based learner feedback.

CASE STUDY / 13

Risks, limitations & responsible claims

A credible quality-assurance portfolio describes not just controls but also how they can fail.

Risk	Mitigation / limitation
False confidence from “pass”	Treat structural success as a narrow result; preserve manual gates.
Quotation matching mistaken for semantics	Use human entailment review or separately validated semantic methods.
A2 label interpreted as certification	Describe alignment to selected aims only.
Synthetic tests misrepresented as model faults	Mark fixtures as injected and report their source honestly.
Audio readiness assumed from a text script	Evaluate real recordings separately.
Independence / affiliation confusion	Identify the work as self-initiated, not a product commissioned by a company.

CASE STUDY / 14

Takeaways & handoff

The strongest outcome is not a claim of perfect language quality, but a structured and inspectable QA workflow.

What this portfolio demonstrates

- Translation of learning-content requirements into a formal style guide and data schema.
- A separation of machine-checkable requirements from human linguistic judgement.
- Evidence-aware documentation of test results, limitations and pending evaluation.
- A practical foundation for controlled prompt evaluation and future listening-task research.

Project files

The companion ZIP contains lesson JSON, two generation prompts, a review prompt, rubric, human review SOP, evaluation report, CSV fixtures/results and an executable Python validator. To reproduce deterministic evidence, run python3 scripts/validate_lessons.py and python3 scripts/run_tests.py from the project root.

References and framework alignment

Council of Europe, Common European Framework of Reference for Languages: Companion Volume (2020). https://www.coe.int/en/web/common-european-framework-reference-languages

NIST, AI Risk Management Framework 1.0 (2023). https://www.nist.gov/itl/ai-risk-management-framework

These frameworks inform review principles; project-specific numbers and release rules are not official standards. Independent portfolio project; no Remba affiliation.