

SEZER MIKAIL SAGIR

AI Evaluation • RLHF • QA Specialist

Bruchsal, Germany (CET) • working remotely | +49 151 29696102 | contact@sezersagir.com
sezersagir.com • linkedin.com/in/sezersagir • github.com/sezersagir

SUMMARY

AI evaluation, RLHF and QA specialist auditing at **senior classification level** on large-scale generative-AI programs via RWS, Volga Partners and Outlier (Expert tier). My focus is the gap between “scores well” and “is actually right”: language that is correct but unnatural, code whose explanation doesn’t match its behaviour, instructions that are quietly dropped. I compare paired outputs against structured rubrics, label edge cases, and turn recurring failures into sharper guidelines — and build the tooling the work needs. Native German and Turkish, English C1, Japanese JLPT N3. Open to freelance, contract and part-time evaluation work.

3,000+ structured evaluations • ~98% QA acceptance • 15–25% lower annotator error • ~20% better cross-team consistency

PROFESSIONAL EXPERIENCE

AI Evaluation, RLHF & QA Specialist

Apr 2025 – Present

RWS, Volga Partners & Outlier (Expert tier) — Independent contractor, Remote • further programs via Alignerr, Scale AI, Agents Only, OneForma • client programs under NDA

- **Direct model evaluation & RLHF.** Paired-output preference rankings with stated reasons; validation of complex code and reasoning chains; TTS audio quality assessment; multilingual naturalness and localisation review in German, Turkish and Japanese.
- **High-volume auditing & QA.** ~120 outputs audited per week within a team of 4–5 auditors; 3,000+ structured evaluations across leading LLMs at a ~98% personal acceptance rate.
- **Annotator guidance & feedback.** Quality oversight for a pod of 13–14 annotators; targeted rubric-based feedback reduced annotator error rates by 15–25%.
- **Guideline & edge-case refinement.** Identified recurring failure modes and edge cases, then closed the rubric ambiguity behind them — improving cross-team grading consistency by roughly 20%.

Sustained full-time, high-volume output while based remotely in Japan (Nov 2025 – Mar 2026) during an intensive Japanese-language immersion period.

Intern, Digital Innovation Lab — CyberForum e.V., Karlsruhe

Jul – Oct 2025

- Synthesised 15+ SME case studies on AI-adoption barriers into a structured strategic analysis; co-organised three startup events within Europe’s largest regional high-tech network (1,400+ member companies).

SELECTED PROJECTS & WRITING

TTS Review Tool — evaluation interface for TTS A/B testing

Sep 2026

Vanilla HTML / CSS / JavaScript — no build step, no dependencies • [live demo](#) • github.com/sezersagir/tts-review-tool

- Keyboard-driven single-page interface, built from the requirements of my own TTS evaluation work and used in live review sessions: CMOS scoring on a –2 to +2 scale, paired A→B or single-track playback at 1×–2×, and multi-select reason tagging for articulation, pacing, naturalness and noise.
- Comment snippets hold terminology steady, autosave restores a session, CSV/TSV export drops straight into a spreadsheet; persistence sits behind a swappable key-value layer. Public demo sanitised for NDA; production build wired to signed client audio URLs.

Real Estate Price Prediction Tool

Apr – Jun 2026

Property valuation for the Karlsruhe/Bruchsal region with a 10-year forecast: an eight-source ensemble over local market and macro data, plus an LLM step that turns free-text property descriptions into a bounded, structured price adjustment. Python, Flask, SQLite, Groq LLM, AWS, Terraform • [GitHub](#)

TikTok Clip Analyzer

Mar 2026

Scores transcribed video segments against a weighted four-criteria rubric — hook strength, standalone value, emotional impact, shareability — returning timestamped, justified clip suggestions. Python, Streamlit, Groq API • [GitHub](#)

Article — “Four failure patterns automated metrics don’t catch”

Sep 2026

Calques that survive a meaning check, code claims the code doesn’t support, negative constraints that quietly disappear, unsupported specificity in grounded summaries. sezersagir.com/blog

EDUCATION

B.Sc. Information Systems — Karlsruhe Institute of Technology (KIT)

from Oct 2026

Abitur — Justus-Knecht-Gymnasium, Karlsruhe

2025

SKILLS

Evaluation RLHF evaluation • quality-rubric review • paired-output comparison • edge-case, error & bias analysis • guideline development • structured feedback • multilingual review (DE / TR / JA / EN)

Technical Python • Java • JavaScript / HTML / CSS • SQL • AWS • Terraform • SageMaker Ground Truth • Scale AI • Labelbox