Anni — Status & Übergabeprotokoll (Stand: 27.08.2025)

Owner: Marisa Langer
System: TranceLate (Self-host MT Worker + Guard „Anni“)
Knoten: MacBook Air (M2/16 GB), lokaler Betrieb

---

1) Executive Summary

Status: Stabil und betriebsbereit für Kern-Use-Case (UI-Texte/Banner/CTA).
Qualität: Platzhalter/HTML/Zahlen werden 1:1 bewahrt; satzbasiert, mit Nach-Cleanup.
Performance (heute): de→en via Guard p50 ≈ **0.18 s** (Promo-Satz), Checks 100 % ok.
Betriebsmodus: CPU, 1 Worker, Concurrency=1, Guard-Batch **off** (maximale Robustheit).
Definition of Done (Kern): /health & /meta ok, TM exact/fuzzy aktiv, Glossar aktiv, Smoke-Tests grün, Invarianten-Checks im Response.

---

2) Architektur (Kurz)

- **Worker (8090):** `mt_server.py` — HF/Opus-MT je Sprachpaar, Gerätewahl über `ANNI_DEVICE`.
- **Guard (8091):** `mt_guard.py` — TM/Glossar, Satz-Split, Segment-Freeze, MT-Call, Restore, Cleanup, Checks.
- **Daten:** `tm.csv`, `glossary.json` (never_translate).
- **Schnittstellen:** `POST /translate` (Guard), `POST /admin/reload`, `GET /meta`, `GET /health`.

---

3) Heute umgesetzte Änderungen (Changelog)

Guard (`mt_guard.py`)

- **Satzweise Verarbeitung** wiederhergestellt; **kein** aggressives Token-Splitten in MT-Segmente.
- **Zahlen-Passthrough**: `PURENUM_RE` wird direkt durchgereicht (kein Maskieren/Restoren nötig).
- **PH/HTML-Freeze** robust: Tolerantes Restore für Marker; Duplikat-Tags werden dedupliziert.
- **PUNC-Maskierung** zunächst deaktiviert; stattdessen **deterministischer Cleanup** (Spacing um `:` und Gedankenstrich `-/–`).
- **Backend-Call gehärtet** (Requests/Retry → später auf urllib getestet, final wieder Requests mit Retry).
- **SINGLE_PH-Fix**: Fallback, falls seltener Split-Edge-Case nur ein einziges PH enthält (Promo-Sätze).
- **Instrumentation**: temporäre Timing-Logs hinzugefügt und **wieder entfernt** (Syntax-Fehler beseitigt).

Worker (`mt_server.py`)

- **Gerätesteuerung** via `ANNI_DEVICE` (cpu|mps|cuda|auto) mit **CPU** als Default (verhindert MPS-Kills auf M2).
- **Thread-sichere Pipeline-Init** (Lock pro Sprachpaar) und **Inference-Semaphor** (gesteuert über `ANNI_MAX_CONCURRENCY`).
- **Start-/Health-Robustheit**: Indentation/Quote-Fixes; Healthcheck stabil.
- **Timing-Log**: Kurzzeitig aktiv zum Debuggen, anschließend entfernt.

Runbooks & Start

- **Stabile Baseline** dokumentiert: `ANNI_DEVICE=cpu`, `MT_WORKERS=1`, `ANNI_MAX_CONCURRENCY=1`, `ANNI_GUARD_BATCH=off`, `ANNI_PREWARM=off`, `ANNI_TORCH_THREADS=1`, `ANNI_TORCH_INTEROP=1`.
- **Prewarm** weiterhin optional (über Guard-Endpoint), aber für Stabilität nicht erforderlich.

---

4) Aktuelle Messwerte (repräsentativ)

- **de→en (Promo-Satz, Guard):** p50 ≈ 0.18 s, p95 ≈ 0.19 s, **Checks: ok** (ph/html/num/paren/len).
- **EN→X Paare:** nach Fixes wieder in Normalbereich (≤1 s) bei Kurzsätzen.
- **Vorherige Ausreißer** (nl→en, it→en p95≈8 s) auf **MPS/Parallelisierung & Mikro-Batching** zurückzuführen → gelöst durch CPU+Semaphor+satzweise.

---

5) Bekannte Stolpersteine & Abhilfe

- **zsh & Here-Docs:** Multi-Zeiler können „hängen“ → in **Einzeilern** arbeiten oder Python-Mini-Snippets separat ausführen.
- **Port-Konflikte:** `address already in use` → `lsof -tiTCP:8090,8091 | xargs kill -9`.
- **Umlaute / Quotes in cURL:** `--data-binary` bevorzugen; JSON vorher testen.
- **Timeouts 6–8 s:** entstehen durch Nebenläufigkeit/Hardware-Tail → Baseline nutzen (CPU/1/1, Batch off).

---

6) Betrieb (One-Step / Copy-Ready)

Start (Baseline, stabil):

```
cd "$HOME/trancelate-onprem"
&& ANNI_DEVICE=cpu ANNI_PREWARM=off ANNI_GUARD_BATCH=off
MT_WORKERS=1 ANNI_MAX_CONCURRENCY=1 ANNI_TORCH_THREADS=1
ANNI_TORCH_INTEROP=1
ANNI_API_KEY=topsecret ./start_local.sh
```

Health & Meta:

```
curl -s http://127.0.0.1:8090/health && echo && curl -s http://
127.0.0.1:8091/meta
```

Smoke (Promo, de→en, über Guard):

```
curl -s -H 'Content-Type: application/json' -H 'X-API-Key: topsecret'
--data-binary '{"source":"de","target":"en","text":"Nur heute: {{COUNT}}
Plätze frei bei <strong>{app}</strong> – 2 Tage gültig!"}'
http://127.0.0.1:8091/translate
```

---

7) Übergabe – Was wurde gefixt (Kurzliste)

- De-/It→En Ausreißer entfernt (CPU-Betrieb, Concurrency-Limit, satzweise Segmente).
- Zahlenerhalt garantiert (Passthrough); AM/PM-Heuristik verbleibt für EN-Quellen.
- PH/HTML 1:1, tolerantes Restore, Dupe-Tag-Dedupe.
- Guard-Cleanup: Spacing um `:` und `-/—`, Len-Ratio Grenzen gemäß Satzlänge.
- TM exact/fuzzy aktiv; `tm.csv` & `glossary.json` Live-Reload über `/admin/reload`.

---

8) Nächste Ausbaustufe (Vorschlag, 24–72 h)

1. **LRU-Memo-Cache im Guard** (pro Sprachpaar, pro Segment-Text) → p50 weiter senken; invalidieren bei `/admin/reload`.
2. **/metrics reaktivieren** (Prometheus-Textformat) mit minimalem Set: uptime, requests_total, errors_total, translate_latency_avg/p50/p95.
3. **Router-Map für Paare** (Vorbereitung für CT2/M2M-Fallbacks); heute bleibt alles auf HF/Opus-MT.
4. **Batch-Runner** `-j 8` als optionaler Lasttest (konservativ steigern, dabei Tail beobachten).
5. **Transcreate-Pfad (separat)** erst nach finaler MT-Stabilisierung einschalten.

---

9) Rollback (falls nötig)

- **Schnellstopp:** `stop_local.sh` ausführen; Ports freimachen; `logs/` prüfen.
- **Configs zurückdrehen:** `ANNI_*` auf Baseline; MPS/CUDA **nicht** aktivieren.
- **Code-Rollback:** letzte funktionierende Fassung aus Repo/Snapshot wiederherstellen.

---

10) Anhänge / Dateien (Kern)

- `mt_server.py`, `mt_guard.py`, `start_local.sh`, `stop_local.sh`, `tm.csv`, `glossary.json`, `logs/`.

Kontakt & Handover:
Alles Nötige ist startbar über die oben genannten Einzeiler. Für die nächste Person: bitte die Baseline-Parameter beibehalten, erst danach (schrittweise) optimieren. Viel Erfolg! 💪