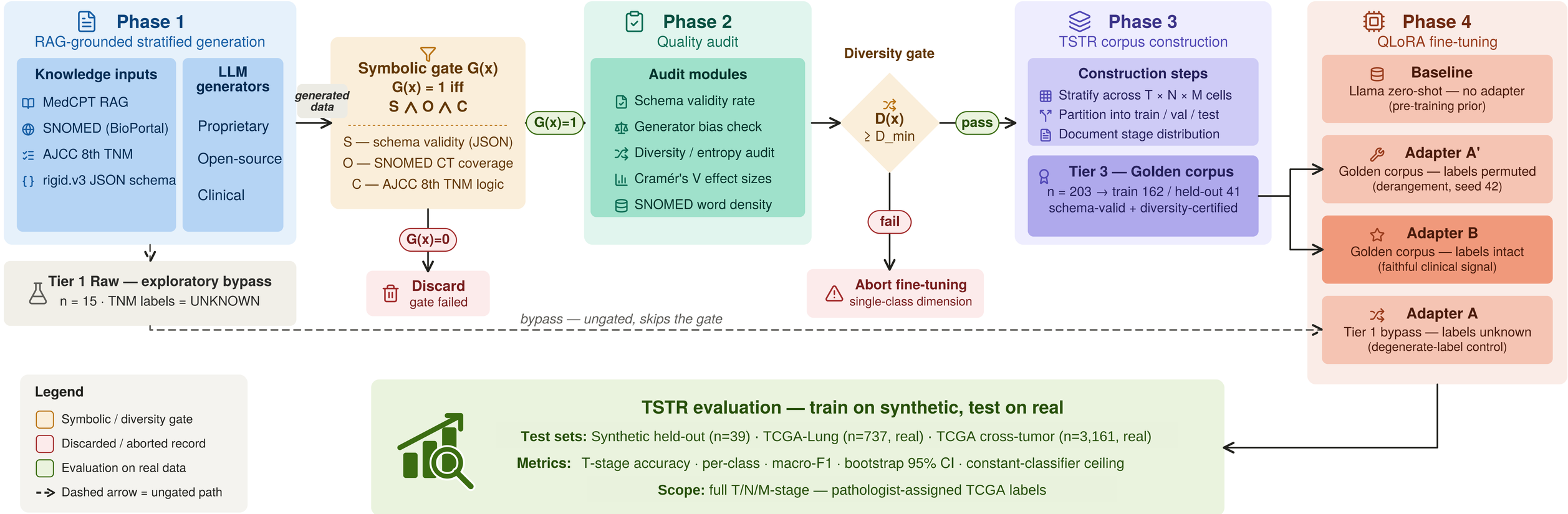

Phase 1
RAG-grounded stratified generation
Knowledge inputs
MedCPT RAG
SNOMED (BioPortal)
AJCC 8th TNM
rigid.v3 JSON schema
LLM generators
Proprietary
Open-source
Clinical
generated data
Symbolic gate G(x)
G(x) = 1 iff
S ∧ O ∧ C
S — schema validity (JSON)
O — SNOMED CT coverage
C — AJCC 8th TNM logic
G(x)=0
Discard
gate failed
G(x)=1
Phase 2
Quality audit
Audit modules
Schema validity rate
Generator bias check
Diversity / entropy audit
Cramér's V effect sizes
SNOMED word density
Diversity gate
D(x)
≥ D_min
pass
fail
Abort fine-tuning
single-class dimension
Phase 3
TSTR corpus construction
Construction steps
Stratify across T × N × M cells
Partition into train / val / test
Document stage distribution
Tier 3 — Golden corpus
n = 203 → train 162 / held-out 41
schema-valid + diversity-certified
Phase 4
QLoRA fine-tuning
Baseline
Llama zero-shot — no adapter
(pre-training prior)
Adapter A'
Golden corpus — labels permuted
(derangement, seed 42)
Adapter B
Golden corpus — labels intact
(faithful clinical signal)
Adapter A
Tier 1 bypass — labels unknown
(degenerate-label control)
Tier 1 Raw — exploratory bypass
n = 15 · TNM labels = UNKNOWN
bypass — ungated, skips the gate
Legend
Symbolic / diversity gate
Discarded / aborted record
Evaluation on real data
Dashed arrow = ungated path
TSTR evaluation — train on synthetic, test on real
Test sets: Synthetic held-out (n=39) · TCGA-Lung (n=737, real) · TCGA cross-tumor (n=3,161, real)
Metrics: T-stage accuracy · per-class · macro-F1 · bootstrap 95% CI · constant-classifier ceiling
Scope: full T/N/M-stage — pathologist-assigned TCGA labels