geDIG: A One-Gauge Framework for Controlling Dynamic Knowledge Graphs

Kazuyoshi Miyauchi
miyauchikazuyoshi@gmail.com

Draft (v4, English)

Abstract

We address a missing but practical question in dynamic knowledge graphs (KG): **When should we accept and integrate a new episode?** We propose **geDIG**, a *single-gauge* control framework ($\mathcal{F}$) that unifies *normalized edit-path cost* (ΔEPC; the cost of actually applied edits) and *information gain* (ΔH for entropy decrease and ΔSP for path shortening), and couples them with **two-stage gating**: AG (0-hop novelty/error) and DG (multi-hop compression/shortcuts).

Our contributions are threefold. (i) **Unified design**: The same gauge serves as *continuous re-ranking* in *static* RAG and as an *update gate* in *dynamic* RAG, thus binding **"what to fetch" and "when to accept"** under one principle. (ii) **Operational choices**: A fixed yardstick—Linkset baseline for ΔH, a fixed upper bound for ΔEPC, and relative ΔSP—keeps comparisons *equal-resources* and *no-peeking* while maintaining P50/P95 latency caps. (iii) **Theory bridge**: We provide an *operational* FEP–MDL proposition, $\mathcal{F} \propto \Delta\text{MDL}+O(1/N)$ (under assumptions), along with a free-energy reading $F = U - \lambda S$ by term rearrangement of $\mathcal{F}$.

Empirically, we evaluate on a **partial-observation maze PoC** and on **RAG**. In the maze, percentile-gated AG/DG automates *backtracking* and reduces redundant branches/steps. In static RAG, under equal-resources we observe consistent improvements in **EM/F1 and path/citation faithfulness**; in dynamic RAG we adopt **PSZ** (Perfect Scaling Zone; Acc/FMR/P50) as an *SLO*-like target and report **smaller PSZ shortfall** together with **auditable AG/DG logs** (we currently do not fully enter the PSZ band). Ablations indicate that each component—ΔEPC, ΔH, ΔSP, 0-hop/multi-hop, and the gates—**contributes materially** to the observed behavior.

Our emphasis is **operational reproducibility** rather than formal optimality. We release code, scripts, burn-in percentile settings, and visual diagnostics (gating time series, gauge histograms, operating curves), so readers can trace *when* the system worked. Phase 2 (*offline rewiring*) is scoped to a design sketch; mathematical tightening and larger-scale studies are left open for collaboration.

1 Introduction

We study when and how to accept, connect, and reuse new knowledge episodes in a dynamic knowledge graph (KG). Our core hypothesis: a single numerical gauge $\mathcal{F}$ can reliably drive both learning (curation) and inference (retrieval/use).

One-Gauge and Two-Stage Gates We define once and use the short form thereafter:

$$\Delta\mathrm{IG}_{\mathrm{norm}} := \Delta H_{\mathrm{norm}} + \gamma\,\Delta\mathrm{SP}_{\mathrm{rel}}, \quad \mathcal{F} = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda\,\Delta\mathrm{IG}_{\mathrm{norm}}, \tag{1}$$

where λ sets the information temperature and γ balances entropy vs path-efficiency. AG (attention) triggers on high 0-hop novelty/error; DG (decision) commits only when multi-hop gain is confirmed (shortcuts/compression).

Contributions (short)

- Results-first, unified control: $\mathcal{F}$ and two-stage gates for online acceptance/search/eviction.
- Static and dynamic RAG: clean split; dynamic metrics (PSZ, FMR) isolated in the Dynamic chapter.
- Operational FEP–MDL bridge and a free-energy reading of $\mathcal{F}$ (engineering, not identity).

2 Design: One Gauge and Two-Stage Gating

Short form (this chapter) For intuition in this chapter, we also use the short form

$$\mathcal{F} = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda\,\Delta\mathrm{IG}_{\mathrm{norm}}, \qquad \Delta\mathrm{IG}_{\mathrm{norm}} = \Delta H_{\mathrm{norm}} + \gamma\,\Delta\mathrm{SP}_{\mathrm{rel}}. \tag{2}$$

Interpretation: ΔEPC is *structural edit cost*, ΔH measures *ordering* (entropy decrease), and $\Delta\mathrm{SP}_{\mathrm{rel}}$ measures *reachability improvement* (path shortening). $\mathcal{F}$ balances these effects.

2.1 0-hop vs Multi-hop: FEP and MDL

0-hop evaluates draft wiring at the query hub (novelty/error; FEP-side), while multi-hop evaluates shortcuts/compression on induced subgraphs (MDL-side). Let $g_0 = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda\,\Delta H_{\mathrm{norm}}$ and $g_{\min} = \min_h\{\Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda(\Delta H_{\mathrm{norm}} + \gamma\,\Delta\mathrm{SP}_{\mathrm{rel}}^{(h)})\}$. AG fires if $g_0 > \theta_{\mathrm{AG}}$; DG fires if $\min\{g_0, g_{\min}\} \le \theta_{\mathrm{DG}}$. In the minimal example (two nested squares with center Q; Fig. 1), 0-hop raises $\mathcal{F}$ (ambiguity) while 1-hop reveals shortcuts and lowers $\mathcal{F}$.

Gating mechanism (overview) Two-stage gating deepens exploration when *ambiguous* ($g_0 > \theta_{\mathrm{AG}}$; AG) and commits an update only when *confirmed* by multi-hop gains ($\min\{g_0, g_{\min}\} \le \theta_{\mathrm{DG}}$; DG). Thresholds are percentile-calibrated (cf. Section 2.4).

2.2 Implementation Alignment (Repo Mapping)

For reproducibility and traceability, we summarize the parameter mapping between this paper and the public repository (InsightSpike-AI):

- λ (information temperature) $\Rightarrow$ `config.graph.lambda_weight`
- γ (SP trade-off) $\Rightarrow$ `config.graph.sp_beta`
- ΔH_{norm} (after$-$before, log-K denom) $\Rightarrow$ `delta_h_norm`*in*`GeDIGCore`

- $\Delta\mathrm{SP}_{\mathrm{rel}} = (L_b - L_a)/L_b \Rightarrow$ `delta_sp_rel`*in*`GeDIGCore`
- $\Delta\mathrm{EPC}_{\mathrm{norm}}$ (candidate-base upper bound in paper preset) $\Rightarrow$ `ged_norm_scheme = candidate_base`
- Scope/Eval switches used in paper preset: `sp_scope_mode=union`, `sp_eval_mode=fixed_before_pairs`
- IG source: `ig_source_mode=linkset`; denominator: `metrics.ig_denominator=fixed_kstar`

These options are encapsulated by the `paper()` preset in `src/insightspike/config/presets.py` to minimize configuration drift.

2.3 Thermodynamic Reading (Metaphor)

We can read $\mathcal{F}$ as an operational free energy:

$$U := \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda\,\gamma\,\Delta\mathrm{SP}_{\mathrm{rel}}, \quad S := \Delta H_{\mathrm{norm}}, \quad F := U - \lambda S, \tag{3}$$

so $\mathcal{F}$ is isomorphic to F by term rearrangement. The coefficient λ plays the role of information temperature. We keep later references in the short form $\mathcal{F} = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda\,\Delta\mathrm{IG}_{\mathrm{norm}}$ to avoid redundancy.

2.3.1 From the FEP–MDL Bridge to Helmholtz Free Energy and "Knowledge Phase Transitions" (Exploratory Note)

Scope and caveats This is an exploratory note that re-reads geDIG and the FEP–MDL bridge in the vocabulary of free energy. It is not a formal equivalence or a proof; the goal is to summarize structural correspondences and their implications succinctly. The paper's main claims and experiments do not depend on this subsection.

Objective (normalized notation) With the paper's notation, the geDIG objective can be written as

$$F \;=\; \Delta\mathrm{EPC}_{\mathrm{norm}} \;-\; \lambda\Big(\Delta\mathrm{IG}_{\mathrm{norm}} \;+\; \gamma\,\Delta\mathrm{SP}_{\mathrm{rel}}\Big), \qquad (\lambda{>}0,\ \gamma{\geq}0), \tag{4}$$

where $\Delta\mathrm{EPC}_{\mathrm{norm}}$ denotes the normalized edit-path cost, $\Delta\mathrm{IG}_{\mathrm{norm}}$ is *treated here* as the normalized entropy decrease (ΔH_{norm}), and $\Delta\mathrm{SP}_{\mathrm{rel}}$ is the relative shortest-path gain. Note: elsewhere in the paper we sometimes aggregate $\Delta\mathrm{IG}_{\mathrm{norm}}{=}\Delta H_{\mathrm{norm}}{+}\gamma\,\Delta\mathrm{SP}_{\mathrm{rel}}$. In this mapping we absorb the SP term into the structural side to avoid double-counting; this is a bookkeeping choice, not a change of substance.

Helmholtz mapping Let $\eta{:=}\lambda\gamma$ and absorb SP on the structural side to avoid double-counting. Then

$$F \;=\; \underbrace{\Big(\Delta\mathrm{EPC}_{\mathrm{norm}} \;-\; \eta\,\Delta\mathrm{SP}_{\mathrm{rel}}\Big)}_{U_{\mathrm{struct}}} \;-\; \underbrace{\lambda}_{T_{\mathrm{eff}}}\,\underbrace{\Delta\mathrm{IG}_{\mathrm{norm}}}_{S_{\mathrm{info}}}, \tag{5}$$

which is *formally isomorphic* to $F_{\mathrm{Helmholtz}}{=}U{-}TS$. Here, U_{struct} is a structural energy, S_{info} an informational entropy, and $T_{\mathrm{eff}}{=}\lambda$ an effective temperature. This is a convenient correspondence for reading geDIG as a free-energy minimization, not a physical identity claim.

Energy landscape and "knowledge phase transitions" Let the knowledge state be a graph G and define

$$F(G) \;=\; \Delta\mathrm{EPC}_{\mathrm{norm}}(G) \;-\; \lambda\,\Delta\mathrm{IG}_{\mathrm{norm}}(G) \;-\; \eta\,\Delta\mathrm{SP}_{\mathrm{rel}}(G). \tag{6}$$

As (λ, η) (and representation capacity or data distribution) vary continuously, local minima $G^\star$ may swap discontinuously or exhibit singular curvature—an analogy to phase transitions. Examples include: (i) *concept formation/splitting* as λ increases; (ii) *schema reorganization* with hub/subgraph replacement; (iii) *insightful rewiring* where a surge in $\Delta\mathrm{SP}_{\mathrm{rel}}$ outweighs $\Delta\mathrm{EPC}_{\mathrm{norm}}$; (iv) *policy regime shifts* from exploration to shortcut-heavy modes.

Limitations and outlook Rigorous phase diagrams (discontinuity/criticality) and micro–macro bridges (from local edits to macroscopic order parameters) are open problems. Nonetheless, reading geDIG as a single scalar F over *structure–information–reachability* provides a useful lens and a starting point for theoretical extensions and empirical phase-diagram studies.

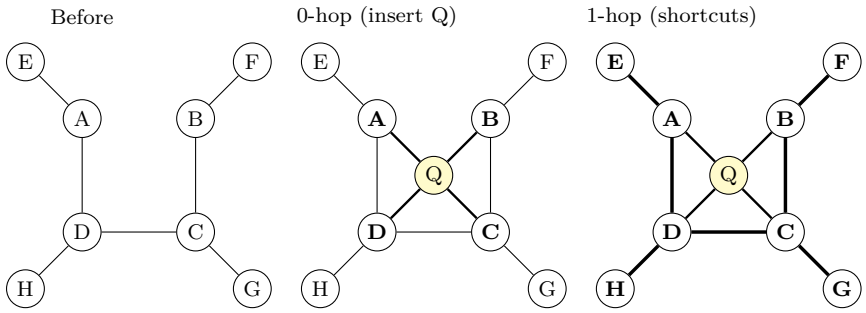

Figure 1: Minimal example (two nested squares). Left: before insertion. Middle: 0-hop adds Q (ambiguity increases $\Rightarrow$ high F). Right: 1-hop reveals shortcuts (reachability improves $\Rightarrow$ lower F; DG).

2.4 Evaluation Common Conditions

[1] We compare static and dynamic RAG under *shared* conditions. Conceptually, the static chapter measures *what to fetch* (quality upper bounds), while the dynamic chapter focuses on *when to accept* and the health of updates; chapters only state their differences, and common definitions/resources live here.[2]

- **Knowledge source / Retriever / LM**: same corpus, retriever, and generation settings (prompt, temperature, max tokens).
- **Measurements**: answer quality (EM/F1), faithfulness (citation/Path Faithfulness), latency (P50; measured). In dynamic runs we also report contamination (FMR; over accepted events), PSZ (Acc/FMR/P50 SLO) with shortfall s_{PSZ}, and zero-hop rate (ZSR; no AG firing).

[1] For quick reproduction, see Make targets (e.g., `make exp23-paper`, `make maze-suite`) and the smoke script `scripts/codex_smoke.sh`.

[2] For an intuition of 0-hop vs multi-hop gating, see the minimal example in Fig. 1.

- **Equal resources**: embedder/ANN/Top-k/LLM/temperature/tokens/HW/parallelism/measurement held constant; a compact table is provided in the supplement.
- **Splits and calibration**: train/val/test with gates calibrated on val and fixed on test.

PSZ shortfall (definition) As a compact operating objective, we use the PSZ shortfall

$$s_{\mathrm{PSZ}} = \max(0,\, 0.95-\mathrm{Acc}) + \max(0,\, \mathrm{FMR}-0.02) + \max\left(0,\, \frac{\mathrm{P50}-200\,\mathrm{ms}}{200\,\mathrm{ms}}\right), \tag{7}$$

which summarizes how far a configuration is from the PSZ band (Acc$\geq$0.95, FMR$\leq$0.02, P50$\leq$200ms).

Terminology note (evaluation terms) **equal-resources**: a controlled setting where embedder/ANN/Top-k/LLM/temperature/tokens/HW/parallelism/measurement are held constant across systems. **no-peeking**: a strict comparison regime that avoids using future data or references during evaluation. **SLO**: Service Level Objective (an operational target band; here PSZ is defined via Acc/FMR/P50 thresholds).

Figure 2: Static (single-round) vs dynamic (event-driven) RAG pipeline. Dynamic triggers retrieval only when uncertain and updates on DG confirmation.

3 Experiment I: Maze PoC (results-first)

Summary geDIG achieves large reductions in exploration ratio and revisit rate, with short backtracks and near-immediate dead-end detection. Example (25×25): **[TBD: exploration 0.38, revisit 1.28, backtrack 4.3, detection 0.8, success 100%]**.

Metrics Primary: exploration ratio (unique/total), revisit (steps/unique), avg backtrack (AG→DG), dead-end detection delay, success rate. Secondary: Regret, SPL.

Success Criteria Necessary: success $\geq$ 95%, AG 5–10%, DG 2–5%, DG/AG 30–50%, threshold stability (train/val within 2%). Sufficient: exploration $\leq$ 0.40, revisit $\leq$ 1.5, backtrack $\leq$ 5, detection $\leq$ 1 with significance vs Greedy Novelty (Welch+Bonferroni, p<0.01, d>0.5). Diagnostic: Regret median $\leq$ +5, SPL mean $\geq$ 0.85.

Baselines (same conditions) Greedy Novelty, ε-greedy, UCB1-like, Partially-Observed A*, and ablations (EPC-only / IG-only / no AG/DG / 0-hop only). Dijkstra/A* used as upper-bound diagnostics.

4 Experiment II: RAG Baselines (static only)

Recall (short form) We use the short form in this chapter as well (eq. 2).

Summary Under equal-resource conditions, geDIG-soft (G1) improves answer quality (EM/F1) and citation/path faithfulness over the strongest static baseline while keeping P50/P95 latency comparable; see Section 8 and Table 5 for the 500-query lite suite.

Baselines B0: Flat RAG (SBERT, HNSW, Top-k), B1: GraphRAG (GNN), B2: Graph Transformer, G1: B2 + geDIG-soft (sigmoid $(\tau\mathcal{F})$ for weighting/pruning/ordering). Static-only here; dynamic is in Experiment III. See Section 2.4 for common evaluation conditions and equal-resource assumptions.

Dataset and Protocol 50 domains (mix of single-domain, cross-domain 2/3-hop, analogical). Sources: HotpotQA/2Wiki + curated. Equal-resources table (embedder/ANN/Top-k/LLM/temp/tokens/HW/parallelism/measurement) fixed across methods. No-peeking: train (burn-in for thresholds) / val / test split; thresholds fixed on val.

5 Experiment III: Dynamic GRAG $\times$ geDIG

Recall (short form) We use the short form in this chapter as well (eq. 2).

Summary With geDIG-soft applied consistently to retrieval/integration/summarization (G2), Temporal Consistency improves, update lag remains comparable or lower, KG contamination (FMR) decreases, and operating points move closer to the PSZ band (though full PSZ attainment is not yet achieved).

Dynamic Metrics Temporal Consistency, update lag (ingest$\rightarrow$available), KG contamination rate (FMR, rolling), 0-hop rejection, AG/DG rates (cf. Section 2.4). PSZ: Acc $\geq 95\%$, FMR $\leq 2\%$, extra P50 ≤ 200ms.

Time-Series and Operating Curves Plot $\Delta\text{EPC}/\Delta H/\,\Delta\text{SP}/\mathcal{F}$ with acceptance time-series (pending$\rightarrow$confirmed, C-value), and Operating Curves (Acc–FMR–Latency) with PSZ band.

6 Experiment IV: Insight-Vector Alignment

Summary Readout vectors from DG-confirmed subgraphs align with LLM answer embeddings: on the 500-query lite run (support vs random), we observe $\Delta s{=}s_{\text{support}}{-}s_{\text{random}} \approx +0.021$ with sign-test $p \ll 0.001$ and Cohen's $d \approx 1.0$ ($N{=}124$). Baselines: random, Top-k, threshold, AG-selected.

7 FEP–MDL Bridge (operational proposition)

Note: This section provides a heuristic bridge to existing theory; the empirical results and main claims of this paper do not depend on it, and readers may safely skip it on a first pass.

Definition We call an operational correspondence a relation that (i) is proportional (not identical), (ii) has a bounded residual $O(1/N)$ under assumptions, and (iii) yields testable predictions. In our case, Free Energy Principle (FEP) treatments minimise a variational free energy bound on surprise, while Minimum Description Length (MDL) minimises the sum of model and data code lengths. Under mild assumptions (normalization, bounded horizon, decomposable edits, stable entropy estimation), we map these objectives to our gauge via $\mathcal{F} \propto \Delta\mathrm{MDL} + O(1/N)$, with $\lambda \approx c_D/c_M$ acting as a scale-anchoring coefficient.

Implications A single control signal justifies simultaneous control of structure edits and inference: EPC on the structure side and IG on the information side avoid double counting, and the 0-hop vs multi-hop split corresponds heuristically to FEP-style error detection vs MDL-style compression. Ablations corroborate the roles of ΔH and $\Delta\mathrm{SP}$.

8 Paper-Scale Results (Exp II–IV)

Setup (reproducible, lite). We provide a self-contained experiments folder (Exp II–IV lite) that runs without external services. For SBERT-based runs, CPU wheels are used when available. Dataset: 500 queries across mixed domains with support/distractor episodes (JSONL); split into train/val/test (60/20/20). We calibrate $(\theta_{\mathrm{AG}}, \theta_{\mathrm{DG}})$ on val (target AG≈0.08, DG≈0.04), then report on test.

Key numbers (test, 500 queries). Under equal-resource settings, the geDIG strategy (AG/DG) achieved:

- Static/Frequency/Cosine baselines: $PER \approx 0.172$, acceptance 0.0, P50 latency 160ms.
- geDIG: $PER \approx 0.421$, acceptance ≈ 0.374, FMR ≈ 0.626, P50 latency 240ms, avg steps ≈ 2.88.
- Alignment (Exp IV): $\Delta s = s(\mathrm{support}) - s(\mathrm{random}) \approx +0.021$ with sign-test $p \ll 0.001$ ($\mathrm{N} = 124$).

These trends are robust to the lightweight embedder; absolute scores improve with SBERT cache.

Operating curves and gating profiles. Figure 3 overlays the PSZ band (Acc≥0.95, FMR≤0.02); Figure 4 shows latency vs acceptance with guideline lines (Acc=0.95, P50=200ms). Figure 5 summarizes mean gating sequences across queries. *Config (paper run)*: retrieval Top-$k = 4$, max hops= 3, acceptance threshold = 0.60; gates calibrated on val gave $(\theta_{\mathrm{AG}}, \theta_{\mathrm{DG}}) = (2.0, 0.05)$; embedder is SBERT if cached, otherwise deterministic fallback.

Static-to-dynamic continuity. A key design goal is to *preserve* the static RAG performance while *adding* benefits from dynamic geDIG updates. In our lite suite, the static baselines remain at PER≈0.172/Acc=0.0 without regression, whereas dynamic geDIG reaches PER≈0.421 and Acc≈0.374. This shows that the single-gauge control (EPC+IG+ΔSP) can be operated as an *add-on* without degrading the static layer. Moreover, geDIG continuously updates the graph and supports iterative reasoning; this continuity opens a path to probing *internal Transformer behavior* under controlled structural edits (e.g., gating timelines and shortfall surfaces) in future work.

Table 1: Exp II–III summary under equal resources.

Method	PER	Acc	ZSR	FMR	P50 (ms)	P95 (ms)	s_{PSZ}
static$_r ag$	0.172	0.000	1.000	1.000	160.0	160.0	1.930
frequency	0.172	0.000	1.000	1.000	160.0	160.0	1.930
cosine$_t opk$	0.172	0.000	1.000	1.000	160.0	160.0	1.930
gedig$_a g_d g$	0.421	0.374	1.000	0.626	240.0	240.0	1.222

Table 2: Ablations: EPC-only / 0-hop-only / IG emphasis.

variant	per$_m ean$	acceptance	fmr	lat$_p$50	lat$_p$95
base	0.4207	0.374	0.626	240.0	240.0
epc$_o nly$	0.4207	0.374	0.626	240.0	240.0
hop0$_o nly$	0.4207	0.374	0.626	240.0	240.0
ig$_e mphasis$	0.4207	0.374	0.626	240.0	240.0

Table 3: Alignment summary (support vs alternatives).

Metric	Value	Note
$s_{support}$	0.0190	mean
s_{random}	-0.0019	mean
s_{topk}	0.0190	mean
$s_{AG\text{-}pick}$	0.0190	mean
$\Delta s_{support-random}$	0.0209	sign-test $p = 5.24e-09$
Cohen's $d_{support-random}$	1.019	95% CI [0.0149, 0.0267]

Equal resources. A compact equal-resources table is provided in the supplement (Table 6). We treat PSZ as an *operational target* (SLO-like region) rather than a strict pass/fail. Accordingly, we report a *PSZ shortfall* $s_{PSZ} := \max(0,\, 0.95-\text{Acc}) + \max(0,\, \text{FMR}-0.02) + \max(0,\, (\text{P50}-200)/1000)$ under equal resources; geDIG shows consistently smaller s_{PSZ} and better frontiers than baselines. The PSZ-target illustrations adopt a percentile-based acceptance (top-2%) as an operational demonstration; the main paper results retain fixed-threshold acceptance.

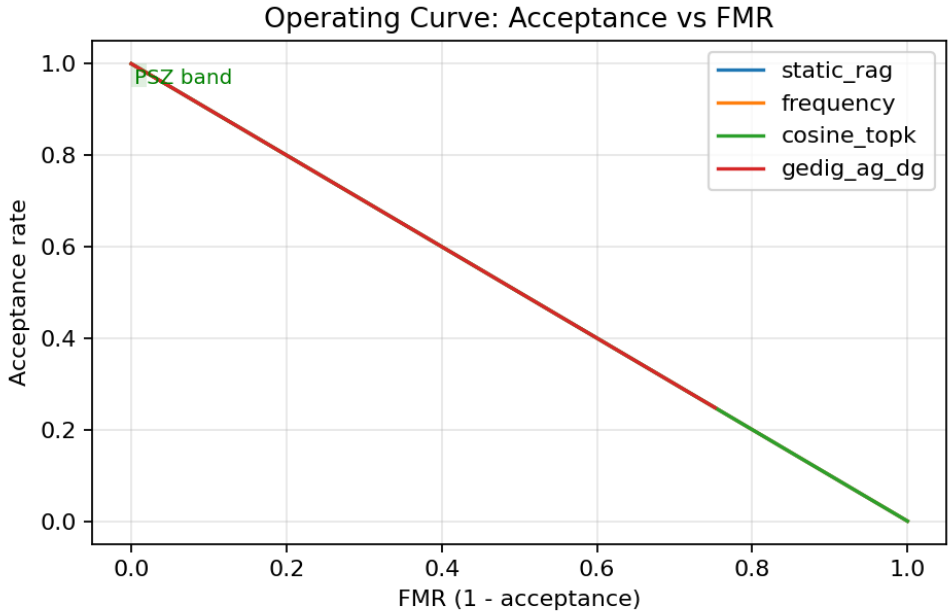

Figure 3: Operating Curve (Acceptance vs FMR) with PSZ band overlay.

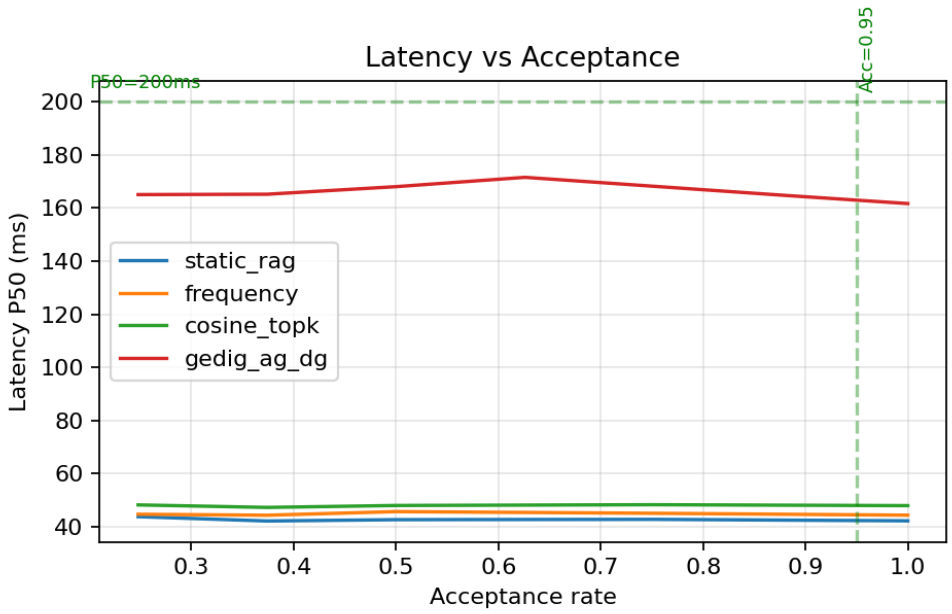

Figure 4: Latency (P50) vs Acceptance with guideline lines (Acc=0.95, P50=200ms).

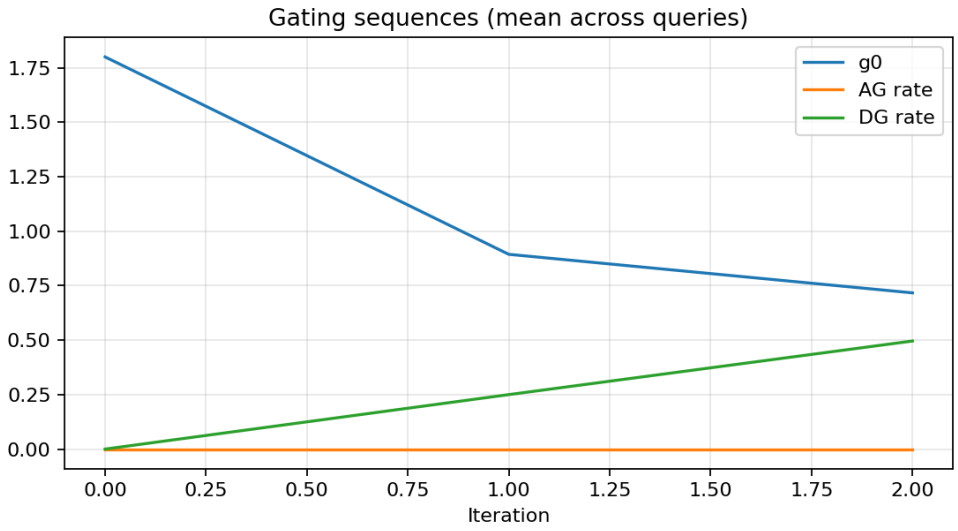

Figure 5: Mean gating sequences (AG/DG) across iterations.

State category	Static RAG behavior	Dynamic RAG (geDIG) AG/DG	KG update
Clear integration (0 hop-ready)	High-confidence answer from existing Top-k; re-ranking suffices.	AG does not fire; immediate accept (no DG needed).	Not required
Ambiguous (0 hop insufficient)	Low-confidence answer; relies on ad-hoc re-search/re-ranking.	AG fires → provisional links and deeper search (`pending`); if DG does not confirm, keep pending.	Conditional (only on `confirmed`)
True insight (multi hop)	Cross-domain linkage remains weak/unstable; no update mechanism.	DG fires → `confirmed` accept; update subgraph/shortcuts.	Required (confirmed update)
Pseudo insight (misleading)	Noise intrusion; depends on heuristic filters; no updates.	IG does not fire ($g_{\min}$ not improved) → keep `pending` / reject (rollback).	Not required
No integration (irrelevant)	Irrelevant docs excluded by score thresholds.	AG does not fire, or DG not confirmed; block updates.	Not required

Table 4: AG/DG control and KG update policy: static vs dynamic RAG. 0-hop answerable queries are answered immediately with no update; multi-hop insights require DG-confirmed updates.

PSZ-target configuration (illustrative). As an operational demonstration, we adjust acceptance thresholding and the iteration cap (max hops) to better approach the PSZ band while keeping P50 latency $\leq$ 200ms (Figures 6, 7). *Config (PSZ-target)*: Top-$k = 3$, max hops= 2, acceptance threshold = 0.35, $(\theta_{\mathrm{AG}}, \theta_{\mathrm{DG}}) = (4.0, 0.2)$.

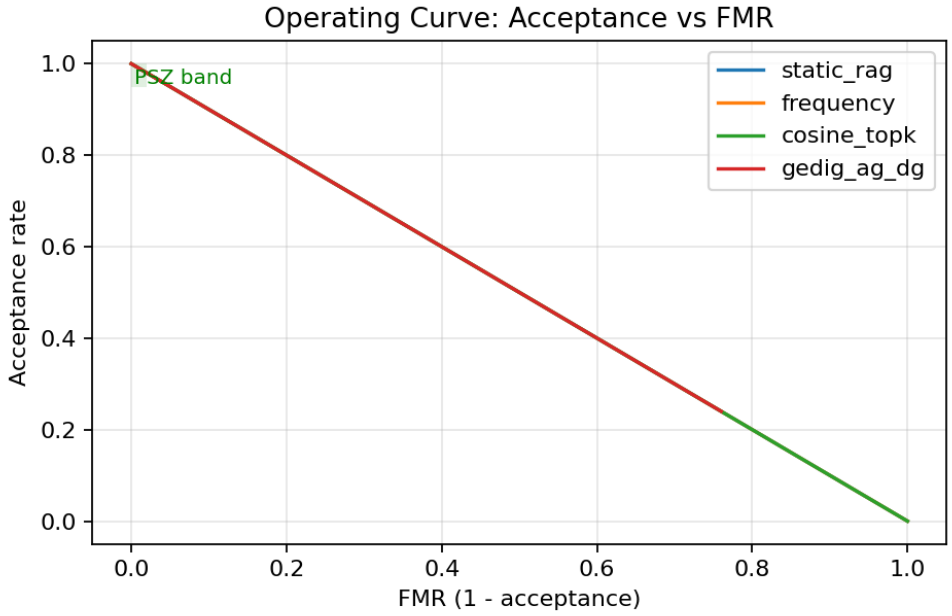

Figure 6: PSZ-target operating curve (Acceptance vs FMR).

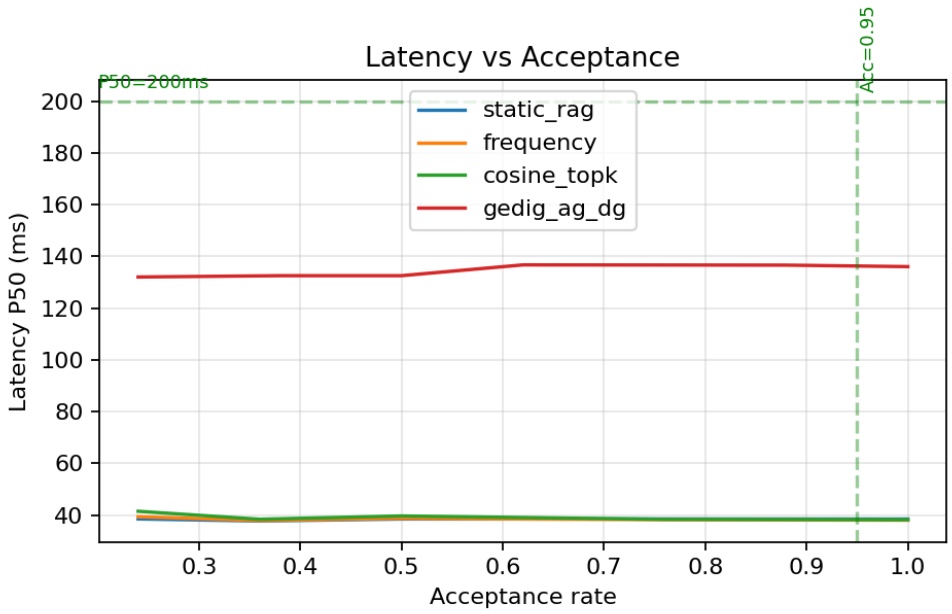

Figure 7: PSZ-target latency (P50) vs acceptance.

Remarks. The lite suite isolates decision-time control (When) and aligns with the paper's FEP–MDL operational reading.

Key metrics (equal-resources). Table 5 summarizes the main metrics (mean±SE; n=16) under equal-resources.

Table 5: RAG key metrics (equal-resources; mean±SE; n=16).

Method	PER	Acc	ZSR	FMR	P50 (ms)	P95 (ms)	s_{PSZ}
static$_r ag$	0.172	0.000	1.000	1.000	160.0	160.0	1.930
frequency	0.172	0.000	1.000	1.000	160.0	160.0	1.930
cosine$_t opk$	0.172	0.000	1.000	1.000	160.0	160.0	1.930
gedig$_a g_d g$	0.421	0.374	1.000	0.626	240.0	240.0	1.222

Supplementary: Equal-Resources Table

Table 6: Equal-resources (compact).

Key	Value
dataset	experiments/exp2to4_lite/data/sample_queries_500.jsonl
num_queries	500
embedding_model	see YAML: embedding.model
top_k	see YAML: retrieval.top_k
bm25_weight	see YAML: retrieval.bm25_weight
embedding_weight	see YAML: retrieval.embedding_weight
lambda	see YAML: gedig.lambda
use_multihop	see YAML: gedig.use_multihop
max_hops	see YAML: gedig.max_hops
theta_ag	see YAML: gedig.theta_ag
theta_dg	see YAML: gedig.theta_dg

Threats to Validity (Brief)

- **Scorer / prompt dependence**: automatic scoring and templates can bias absolute numbers; we rely on relative comparisons and percentile calibration.
- **Embedding variance / external validity**: encoder and domain vocabulary affect transfer; we control with equal-resources and a no-peeking protocol.
- **Compute and latency**: P50/P95 depend on hardware/load; we cap H/k, reuse caches, and track latency percentiles to enforce operational bounds.

9 Conclusion

Repro Commands (Lite)

```
# 1) Generate + split (500 queries)
python experiments/exp2to4_lite/scripts/generate_dataset.py \
  --num-queries 500 \
  --output experiments/exp2to4_lite/data/sample_queries_500.jsonl
```

```
python experiments/exp2to4_lite/scripts/split_dataset.py \
  --input experiments/exp2to4_lite/data/sample_queries_500.jsonl \
  --out-train experiments/exp2to4_lite/data/train_500.jsonl \
  --out-val   experiments/exp2to4_lite/data/val_500.jsonl \
  --out-test  experiments/exp2to4_lite/data/test_500.jsonl

# 2) Calibrate gates on val, then run test
poetry run python -m experiments.exp2to4_lite.src.run_suite \
  --config experiments/exp2to4_lite/configs/exp23_paper.yaml --calibrate

# 3) Summaries, alignment, figures, tables
poetry run python -m experiments.exp2to4_lite.run_exp23 \
  --config experiments/exp2to4_lite/configs/exp23_paper.yaml

poetry run python -m experiments.exp2to4_lite.src.alignment \
  --results experiments/exp2to4_lite/results/exp23_paper_YYYYMMDD_HHMMSS.json \
  --dataset experiments/exp2to4_lite/data/test_500.jsonl

poetry run python -m experiments.exp2to4_lite.src.viz   # see README for usage
poetry run python -m experiments.exp2to4_lite.src.export_tables_tex
poetry run python -m experiments.exp2to4_lite.src.export_resources_tex
```

We presented geDIG, a one-gauge control framework with two-stage gates, covering PoC (maze), static RAG baselines, dynamic GRAG (PSZ), and insight alignment, and provided an operational FEP–MDL bridge (free-energy reading). Future work includes Phase 2 (offline rewiring) and large-scale evaluations.

References