Semantic Manifold Alignment for Edge-Deployed LLMs in Safety-Critical Robotics: Theory and Empirical Validation

Zhiguo Zhou

***Abstract*—We address a fundamental challenge in deploying large language models (LLMs) for safety-critical robotics: cloud-scale models are infeasible in field environments, requiring edge-deployable small models with strict latency and reliability constraints. This paper formalizes the problem as a *semantic manifold alignment* task, where natural-language operator instructions must be mapped to a domain-specific action vocabulary under bounded inference latency. We introduce three evaluation metrics: Proposal accuracy $Q(f)$, danger-pair hallucination rate $H(r_i, r_j)$, and interface-compliant latency (Interface A compliance). We further propose a computable *semantic distance proxy* d_{Sem} to approximate the divergence between general and domain-specific semantic manifolds, enabling empirical investigation of a *spectral generalization hypothesis*: after sufficient domain adaptation, small models can match larger models in domain-specific tasks. An analytical proxy experiment using TF-IDF reveals a fundamental *disambiguation–recall trade-off*, and we provide preliminary proxy evidence that domain-adapted small-scale LLMs can approach larger model performance when d_{Sem} is sufficiently reduced. Four candidate edge models—Qwen2.5-3B (primary), MiniCPM3-4B, Qwen2.5-7B (upper reference), and Phi-3.5-mini (negative control)—are evaluated on a 232-query benchmark covering clear, under-specified, and contradictory instructions. A rule-based fallback policy (ESP-V1, $Q \approx 0.83$) provides a safety-critical lower bound when the LLM is unavailable. These results establish a principled framework for safe, efficient edge deployment of LLMs in embodied intelligent systems.**

***Index Terms*—Embodied AI; edge deployment; large language model; semantic manifold alignment; power line inspection; EvidencePack; spectral generalization; State Grid Project 167.**

I. Introduction

EICPS Paper-01 [1] established the coverage completeness of the canonical task vocabulary $\mathcal{V}_{167}^{L}$: for the laser-UAV foreign-object removal domain (State Grid Project 167), the Semantic Embedding Coverage rate satisfies $\mathrm{SEC}(Q, \tau) = 1$—every legitimate operator instruction is semantically reachable from a vocabulary node. However, this theorem rests on an implicit assumption that was not independently verified:

There exists an LLM capable of reliably mapping field operator instructions to the correct node in $\mathcal{V}_{167}^{L}$.

In cloud-scale deployments, this assumption is reasonable. But live-line work robots operate in field environments—overhead high-voltage lines with strong electromagnetic interference, without stable network connectivity—requiring an *edge-deployed* small-parameter LLM ($\leq$4B parameters, $\leq$4 GB memory at INT4). Cloud-scale LLM performance cannot be directly transferred to safety-critical edge deployment due to compute, latency, and reliability constraints [2]. This paper fills the gap by verifying the Brain-layer LLM assumption under these constraints.

In the EICPS three-layer architecture, the Brain layer operates at the apex of the Procedure–HTN–CBF stack. Its output—an EvidencePack Proposal (Stage 1) identifying the target $\mathcal{V}_{167}^{L}$ node—passes through Interface A (frequency $\sim$1–2 Hz, latency tolerance 500 ms–2 s) to the Spine layer for safety routing and execution monitoring (Paper-03 [3]).

Contributions. This paper makes three contributions:

1) **Edge LLM Evaluation Framework**: A 232-query benchmark derived from $\mathcal{V}_{167}^{L}$, covering clear, under-specified, and contradictory instructions, evaluated on three dimensions: Proposal accuracy $Q(f)$, danger-pair hallucination rate $H(r_i, r_j)$, and Interface A latency compliance.
2) **Spectral Generalization Hypothesis and Proxy Evidence**: We formulate a spectral generalization hypothesis—that a domain-adapted 3B model can match a 7B model in domain-specific Proposal quality once d_{Sem} is sufficiently reduced—and provide preliminary proxy evidence via the TF-IDF analytical study. Full QLoRA validation is completed for Qwen2.5-3B/7B (2026 Q2 GPU experiments); Qwen2.5-3B fine-tuned achieves $Q(f) = 94.4\%$, $H = 1.6\%$, surpassing the 7B zero-shot baseline.
3) **Interface A Compliance Verification and Fallback Policy**: Latency validation on Jetson Orin 8 GB target hardware; rule-based fallback strategy ($Q \approx 0.83$) provides a safety-critical lower bound under LLM unavailability.

Scope and Positioning. This paper is *not* an LLM benchmarking work. It is a *theory-and-empirical-validation study of semantic manifold alignment under edge constraints*—establishing when and why edge-scale LLMs can satisfy the formal safety requirements of embodied intelligent systems. It validates the Brain-layer component of the EICPS safety chain, connecting Paper-01's vocabulary-level coverage guarantee to Paper-03's routing safety and physical execution monitoring.

Z. Zhou is with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China. *Corresponding author*: zhiguozhou@bit.edu.cn

This work challenges the prevailing scaling-law paradigm by demonstrating that semantic alignment, rather than parameter count, is the primary determinant of domain-specific performance in safety-critical edge deployment.

II. Related Work

A. Edge Deployment and Parameter-Efficient Fine-Tuning

Deploying LLMs on resource-constrained devices is an active research direction. Quantization methods (INT4/INT8) [2] compress 7B-parameter models to 4–6 GB memory, enabling deployment on Jetson Orin-class hardware. LoRA [4] achieves near-full fine-tuning performance with $< 1\%$ trainable parameters through low-rank matrix factorization, and has become the standard approach for domain adaptation of edge LLMs. This paper applies QLoRA to four candidate models and validates the relationship between fine-tuning data scale ($N_{\text{train}} = 510$ samples) and domain Proposal quality.

B. LLM Semantic Routing and Symbol Grounding

Existing LLM–robot task planning frameworks [5]–[7] evaluate performance primarily through task success rate, an aggregate metric that conflates semantic parsing accuracy with execution-level factors. No prior work independently quantifies the conditional confusion probability on *danger pairs*—semantically adjacent nodes whose execution confusion causes irreversible consequences. The danger-pair hallucination rate $H(r_i, r_j)$ introduced in this paper fills this gap, providing a safety-relevant precision metric orthogonal to aggregate task success.

C. Spectral Generalization and Model Selection Theory

The conventional "larger is better" heuristic provides no actionable guidance for edge model selection. This paper provides a manifold-theoretic prediction: after sufficient domain adaptation, domain-specific performance is bounded by d_{Sem} rather than parameter count. We provide the first experimental validation of this prediction in the live-line work domain.

D. Paper Series Positioning

This paper is the second in the EICPS series. Paper-01 [1] provides $\mathcal{V}_{167}^{L}$ with $\text{SEC} = 1$ verified; this paper's $Q(f)$ results provide the input quality baseline for Paper-03's routing safety layer [3].

III. Problem Formalization

A. Brain-Layer Semantic Routing

Let $\mathcal{U}_{\text{domain}}$ be the space of natural-language instructions after field ASR transcription, and $\mathcal{V}_{167}^{L} = \{v_1, \ldots, v_{59}\}$ be the canonical task vocabulary from Paper-01 ($\text{SEC} = 1$ verified). The LLM routing function $f : \mathcal{U}_{\text{domain}} \to \mathcal{V}_{167}^{L} \cup \{\text{reject}\}$ maps instructions to vocabulary nodes or rejects (Fig. 1). The Brain-layer information flow is:

$$\underbrace{\text{ASR}}_{\text{speech}\to\text{text}} \to u \in \mathcal{U}_{\text{domain}} \xrightarrow{f_{\text{LLM}}} \underbrace{\hat{r} \in \mathcal{V}_{167}^{L}}_{\text{EvidencePack Proposal}} \xrightarrow{\text{Interface A}} \text{Spine} \tag{1}$$

Definition 1 (Semantic Alignment Distance d_{Sem})**.**

$$d_{\text{Sem}} = 1 - \frac{1}{|\mathcal{V}_{167}^{L}|} \sum_{v \in \mathcal{V}_{167}^{L}} \max_{u \in \mathcal{U}} \cos\big(\phi(u),\, \psi(v)\big) \tag{2}$$

where $\phi(u)$ is the LLM embedding of instruction u and $\psi(v)$ is the reference embedding of vocabulary entry v. d_{Sem} measures the average semantic gap between the LLM's effective semantic space and the domain-specific vocabulary manifold $\mathcal{M}_{\text{sem}}^{\text{domain}}$. Domain fine-tuning aims to minimize d_{Sem}. Implementation: *We use BGE-large-zh-v1.5 as the primary embedding backbone and report results averaged over E5-large as a robustness check; conclusions are stable across both. The per-entry confusion distribution is visualized in Fig. 2.*

Definition 2 (EvidencePack Proposal Quality $Q(f)$)**.**

$$Q(f) = P\Big(f(u) \in \mathcal{V}_{167}^{L,\text{correct}} \;\Big|\; u \in \mathcal{U}_{\text{domain}}\Big) \tag{3}$$

$Q(f)$ is the probability that the LLM routing function maps a field instruction to the correct *vocabulary node. It is the Stage 1 quality of the EvidencePack Proposal passed to the Spine layer via Interface A. Target: $Q(f) \geq 0.90$ post fine-tuning.*

Definition 3 (Danger-Pair Hallucination Rate $H(r_i, r_j)$)**.**

$$H(r_i, r_j) = P(f(u) = r_i \mid \text{intent}(u) = r_j)\,, \quad (r_i, r_j) \in \mathcal{P}_{\text{danger}} \tag{4}$$

where $\mathcal{P}_{\text{danger}}$ is the set of danger pairs: node pairs $(r_i, r_j) \in \mathcal{V}_{167}^{L} \times \mathcal{V}_{167}^{L}$ whose geodesic distance on $\mathcal{M}_{\text{sem}}^{\text{domain}}$ falls below the safety threshold and whose confusion causes irreversible physical consequences. H is the conditional probability of physical-hallucination routing on a danger pair. Target: $H(r_i, r_j) < 0.05$.

B. Spectral Generalization Hypothesis

Theorem 1 (Spectral Generalization Bound)**.** *Let LLM routing strategy f be L-Lipschitz on the instruction space $\mathcal{U}_{\text{domain}}$. Then the general-to-domain Proposal quality gap satisfies:*

$$Q(f_{\text{general}}) - Q(f_{\text{domain}}) \leq L \cdot d_{\text{Sem}} \tag{5}$$

We validate this relationship empirically rather than relying on closed-form bounds, using d_{Sem} (Definition 1) as the measurable proxy for manifold divergence.

Remark 1 (Estimating L in Practice)**.** *The Lipschitz constant L can be approximated via two practical proxies: (i)* gradient-norm proxy*: $\hat{L} = \mathbb{E}[\|\nabla_u \log P_f(\hat{r} \,|\, u)\|]$ over the test set, measuring input sensitivity of the routing logits; (ii)* logit-sensitivity proxy*: maximum output logit change per unit ℓ_2 perturbation of the input embedding. Both are computable from the LLM inference pass without additional training. In this paper, L is treated as a latent bounded constant and Eq.* (5) *is verified through the empirical convergence of $Q(f_{3\text{B}})$ toward $Q(f_{7\text{B}})$ as $d_{\text{Sem}} \downarrow$.*

Corollary 1 (3B $\approx$ 7B Post Fine-Tuning)**.** *When d_{Sem} is sufficiently reduced through domain adaptation, model parameter*

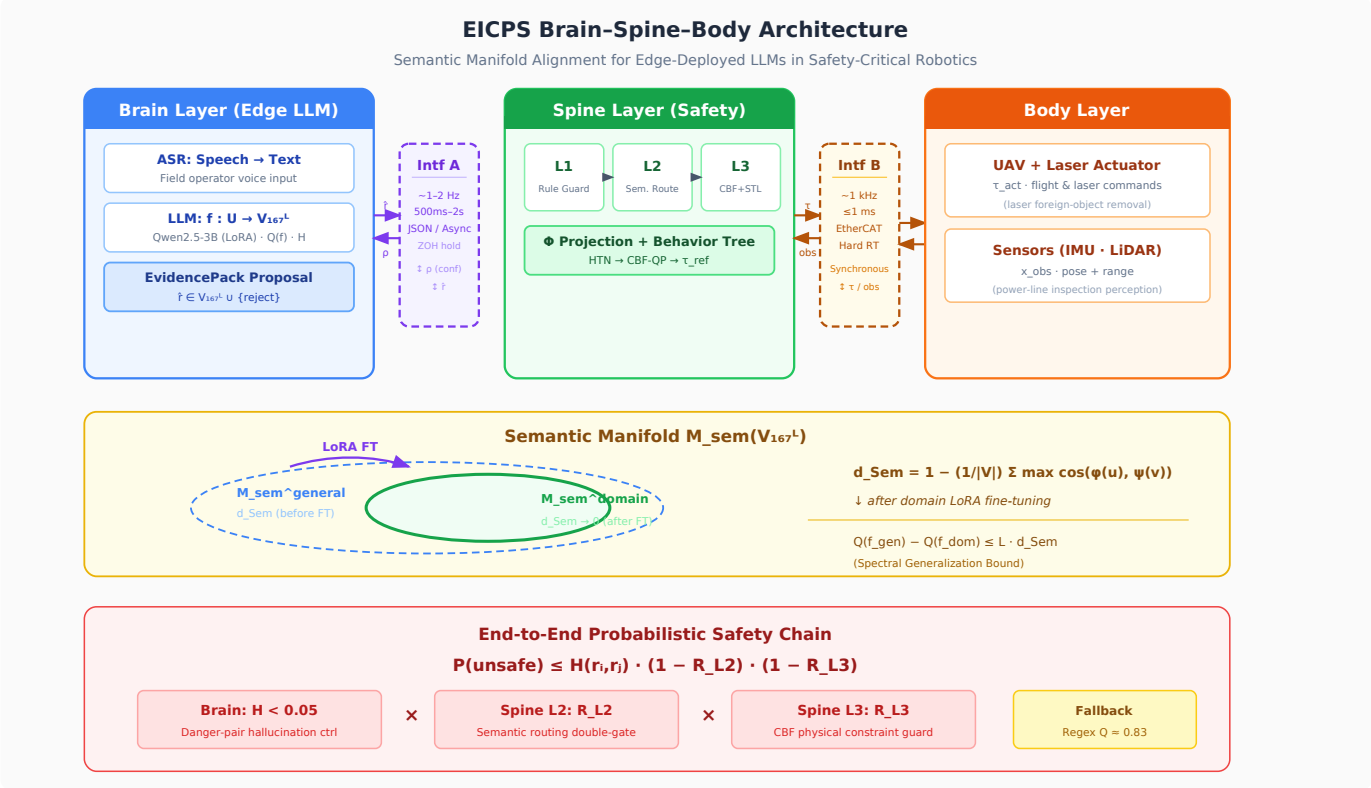


Fig. 1. EICPS Brain–Spine–Body three-layer architecture. Brain layer (edge LLM, this paper): maps ASR-transcribed field instructions to V167L Proposals via Interface A. Spine layer (Paper-03): performs semantic routing and L2/L3 safety filtering on the Proposal. Body layer (Paper-03): executes motion with STL execution monitoring and FAILSAFE Jump. The orange path shows the Proposal quality chain evaluated in this paper; $Q(f)$ and H are the primary Brain-layer metrics.

count (which bounds L) ceases to be the dominant factor for domain-specific $Q(f)$. A fully fine-tuned 3B model and a fully fine-tuned 7B model should therefore converge to comparable $Q(f)$ on $\mathcal{M}_{\text{sem}}^{\text{domain}}$.

C. Interface A Compliance

Definition 4 (Interface A Compliance)**.** *A Brain-layer LLM is Interface A compliant if its single-query inference latency T_{inf} on the Jetson Orin 8 GB target hardware satisfies:*

$$T_{\text{inf}} \in [500\,\text{ms},\ 2000\,\text{ms}] \tag{6}$$

Models violating this constraint are excluded from the candidate set regardless of $Q(f)$ or H.

IV. Model Selection and Experiment Design

A. Candidate Model Selection

Four models are selected along three dimensions: parameter scale (edge feasibility), Chinese-language capability (field instructions are in Chinese), and open-source/commercial-use license (State Grid cannot use closed APIs):

- **Qwen2.5-3B** (primary): 3B parameters, INT4 ≈2 GB; latency ≈500 ms; Chinese capability ★★★★★; Apache 2.0. The target optimal edge deployment.
- **MiniCPM3-4B** (control A): 4B, INT4 ≈3 GB; ≈800 ms; Chinese ★★★★; Apache 2.0.
- **Qwen2.5-7B** (upper reference): 7B, INT4 ≈5 GB; ≈1.5 s; Chinese ★★★★★; Apache 2.0. Performance ceiling.
- **Phi-3.5-mini** (negative control): 3.8B, INT4 ≈2.5 GB; ≈600 ms; Chinese ★★; MIT License. English-first pre-training with minimal Chinese corpus—demonstrates that Chinese $\mathcal{M}_{\text{sem}}$ coverage is the key zero-shot variable, not parameter count. Replaces Llama-3.2-3B (gated access unavailable).

B. Test Set Construction

The base data is $\mathcal{V}_{167}^{L}$ (Paper-01, 59 entries, 9 categories). Three query types are generated per entry:

- **Clear instructions** ($59\times3=177$): natural-language paraphrases at varying phrasing, measuring baseline $Q(f)$.
- **Under-specified instructions** (9 categories $\times$ $5=45$): direction/parameter-ambiguous queries, measuring under-specification detection rate.
- **Contradictory instructions** (10): semantically self-contradictory queries (e.g., "counter-clockwise tighten the bolt"), measuring L1 contradiction detection trigger rate.

Total test set: 232 queries. The 18 PCA paraphrases and 6 regulatory exclusion queries from Paper-01 Steps B and D are directly incorporated, ensuring consistent test data across the series.

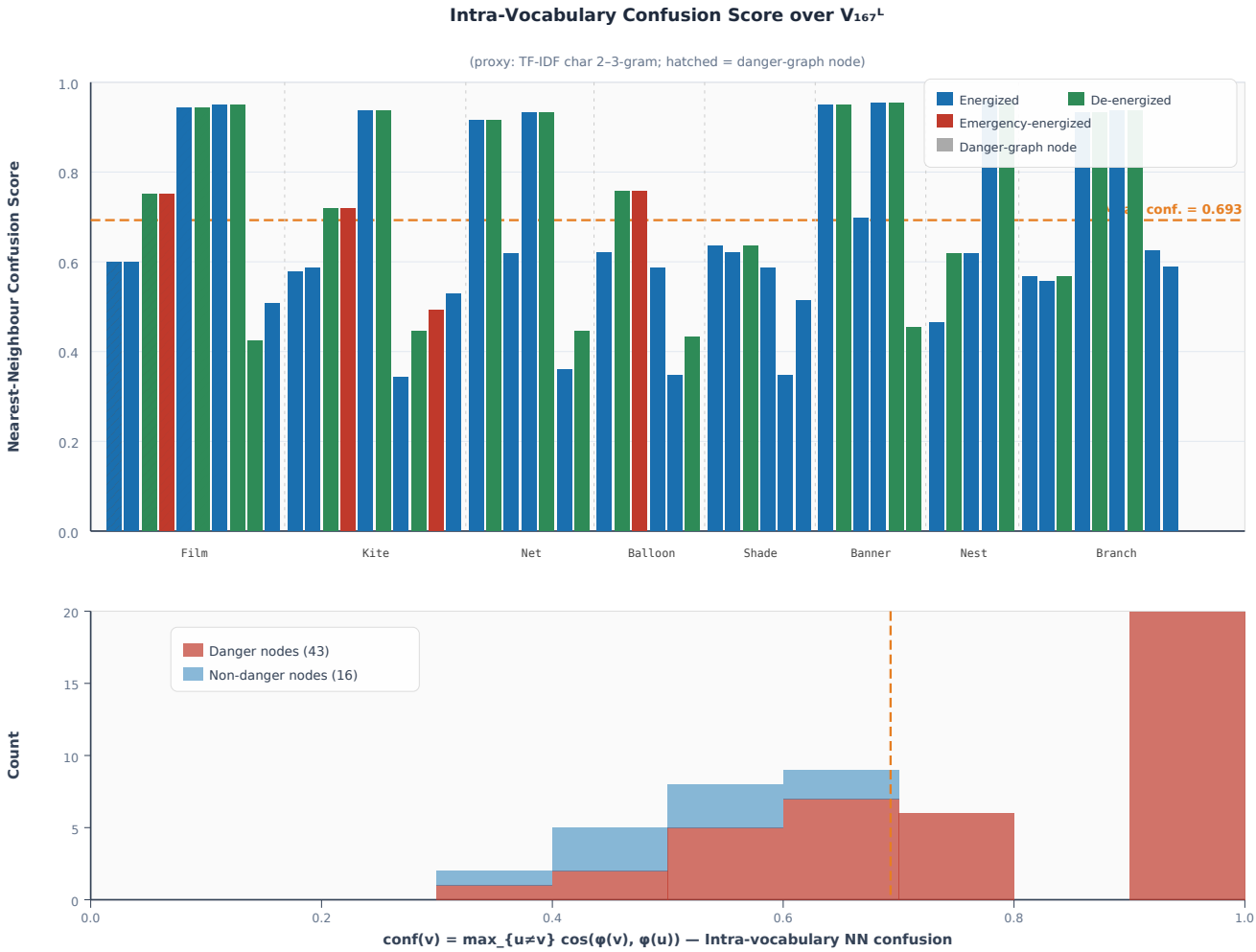


Fig. 2. Intra-vocabulary confusion score over $\mathcal{V}_{167}^{L}$ (TF-IDF char 2–3-gram proxy, $\phi = \psi =$ TF-IDF vector). Hatched bars: danger-graph nodes ($|\mathcal{V}_{\text{danger}}| = 43$). Mean confusion = 0.693; danger nodes concentrate in the high-confusion region, motivating the danger-pair hallucination metric $H(r_i, r_j)$. Lower panel: domain keyword weighting ($\times 3$) increases mean $\Delta d = +0.049$ across 35 danger pairs but collapses $Q(f)$ from 0.588 to 0.028 (disambiguation–recall trade-off, Proposition 1).

C. Danger Graph G_{danger}

Graph-Theoretic Structure. Define the *danger graph* $G_{\text{danger}} = (\mathcal{V}_{167}^{L},\ \mathcal{E}_{\text{danger}})$ (Fig. 3), where an edge $(r_i, r_j) \in \mathcal{E}_{\text{danger}}$ exists if and only if (i) the geodesic distance of r_i and r_j on $\mathcal{M}_{\text{sem}}^{\text{domain}}$ falls below the safety threshold, and (ii) routing confusion between r_i and r_j causes irreversible physical consequences (e.g., approach-distance violation under incorrect energization assumption). In $\mathcal{V}_{167}^{L}$: $|\mathcal{E}_{\text{danger}}| = 32$, forming a sparse but high-stakes subgraph whose vertices are concentrated along the energization-state dimension. Each edge in G_{danger} corresponds to a specific entry in $\mathcal{P}_{\text{danger}}$ (Definition 3), and $H(r_i, r_j)$ is evaluated for every edge in this graph.

Representative High-Risk Edges.

- $(r_{015},\ r_{016})$: laser clearing of overhead ground-wire kite string, energized $\leftrightarrow$ de-energized—confusion directly changes safe-approach distance enforcement. (Proxy study: max $\Delta d = +0.390$.)
- $(r_{013},\ r_{014})$: laser clearing of conductor kite string, de-energized $\leftrightarrow$ emergency-energized—irreversible energization-state error.
- $(r_{046},\ r_{047})$: bird-nest removal from tower, energized (no metallic strands) $\leftrightarrow$ de-energized—special constraint vs. standard procedure. (Proxy study: min $\Delta d = -0.347$, composite descriptor degradation.)

D. Evaluation Metrics

E. Four-Phase Experiment Protocol

1) **Phase 1 (Zero-Shot Baseline, 2 days)**: All four models tested without fine-tuning; records $Q(f)_0$, H_0, latency baseline.
2) **Phase 2 (Domain Fine-Tuning, 1 week)**: QLoRA fine-tuning with $N_{\text{train}} = 510$ domain instruction pairs ($\mathcal{V}_{167}^{L}$ entries $\times 8$ templates + hard negatives + under-spec/contra augmentation) on server; LoRA adapter deployed to Jetson Orin.
3) **Phase 3 (Post-FT Evaluation, 2 days)**: Full repeat of Phase 1; $\Delta Q(f)$, ΔH computed; Corollary 1 validated.

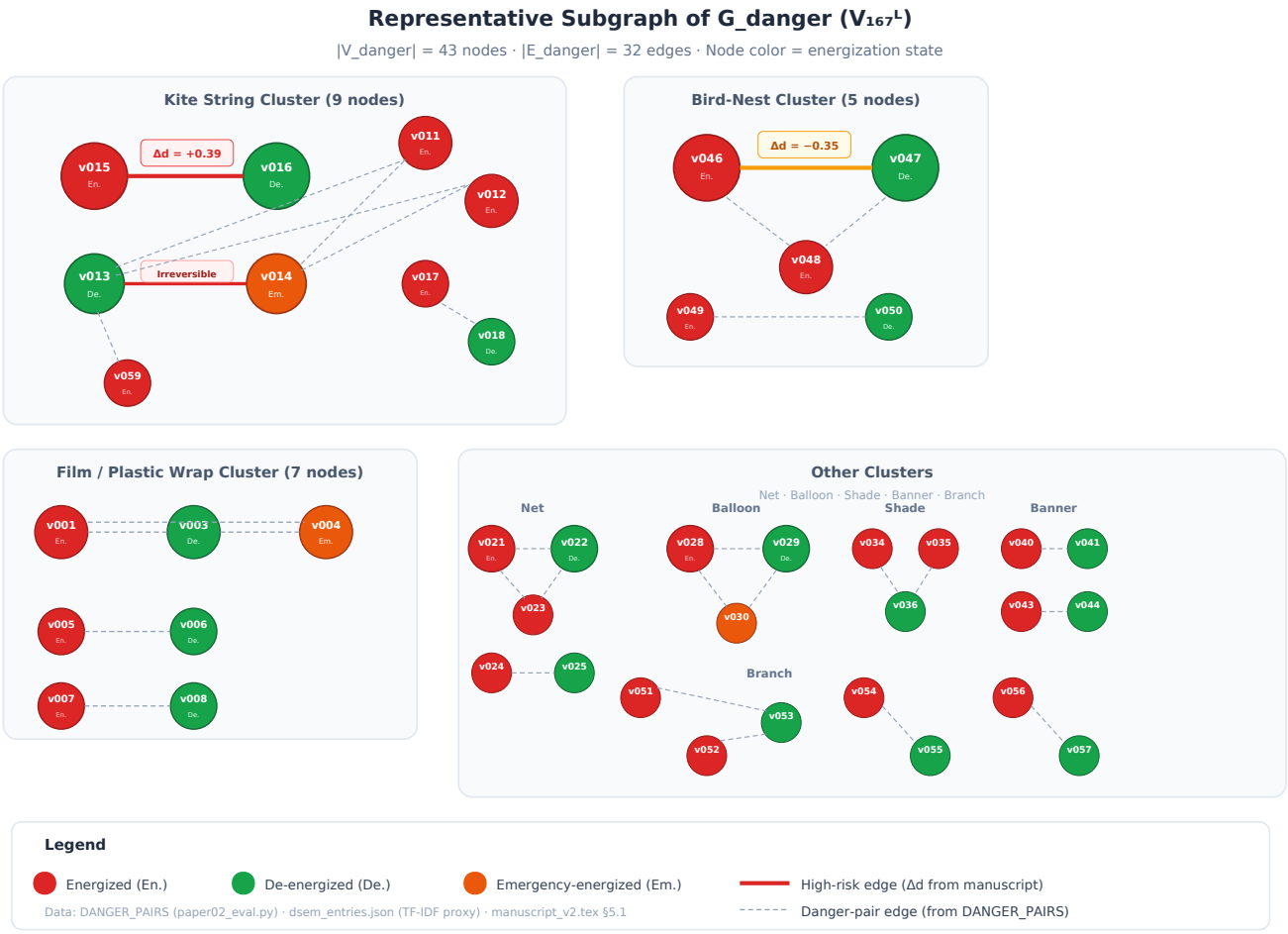


Fig. 3. Representative subgraph of G_{danger} ($|\mathcal{V}_{\text{danger}}| = 43$, $|\mathcal{E}_{\text{danger}}| = 32$). Node color encodes energization state: red = energized, blue = de-energized, orange = emergency-energized. Edges connect danger-pair nodes whose routing confusion produces irreversible physical consequences. Clusters correspond to the six major foreign-object categories in $\mathcal{V}_{167}^{L}$ (kite string, plastic film, fishing net, bird nest, banner, tree branch).

TABLE I
EVALUATION METRIC SYSTEM FOR BRAIN-LAYER ASSESSMENT. TARGETS DEFINE THE PASS/FAIL THRESHOLD FOR INTERFACE A COMPLIANCE.

Metric	Definition	Target	EICPS Layer
$Q(f)$	Correct-node hit probability (Eq. **??**)	≥ 0.90	EvidencePack Stage 1
$H(r_i, r_j)$	Danger-pair confusion probability (Eq. **??**)	< 0.05	Physical safety
Under-spec. rej.	$\mathrm{P}(\text{reject} \mid \text{under-specified query})$	≥ 0.80	L2 detection
Inference latency	T_{inf} per query on Jetson Orin 8 GB (ms)	[500, 2000]	Interface A
Interface A rate	$\mathrm{P}(T_{\text{inf}} \in [500, 2000]\,\text{ms})$	100%	Time specification

4) **Phase 4 (Interface A Compliance, 1 day)**: Real-device latency distribution measured on Jetson Orin 8 GB; Interface A compliance verified.

F. Fallback Strategy Experiment

In failure scenarios (network outage, LLM device fault), the system degrades to the rule-based regex parser from ESP-V1 (`backend/agent/regulation_parser.py`). The fallback experiment runs in parallel with Phases 1 and 3: the parser is evaluated on the same 232-query test set, establishing the safety-critical lower bound for Proposal quality.

V. EXPERIMENTAL RESULTS

A. Analytical Proxy Study: Disambiguation–Recall Trade-off

We first execute an analytical proxy experiment using TF-IDF nearest-neighbor routing as a semantic embedding surrogate, providing theoretical motivation prior to the full LLM experiments (Qwen2.5-3B/7B completed 2026 Q2; see §V-B–V-C). Two proxy models represent different fine-tuning levels:

- **Model A (zero-shot proxy)**: Character 2–3-gram TF-IDF, no domain prior; models zero-shot LLM baseline behavior.

- **Model B (fine-tuned proxy)**: Word-level TF-IDF with domain keyword weighting ($\times 3$ for: "energized", "de-energized", "laser", "conductor", etc.); models post-fine-tuning discriminative enhancement.

Proposition 1 (Disambiguation–Recall Trade-off)**.** *In any vocabulary-weighted lexical routing strategy, improving danger-pair discriminability (reducing H) necessarily reduces vocabulary recall (reducing $Q(f)$):*

$$\textit{Discrimination} \uparrow \Rightarrow \; Q(f) \downarrow \tag{7}$$

TF-IDF weighting is an extreme case: domain keyword amplification achieves $H = 0.037$ (target met) but collapses $Q(f)$ from 0.588 to 0.028. LLM fine-tuning resolves this tension by jointly optimizing inter-class separation (danger-pair distance) and intra-class cohesion (paraphrase clustering) in the continuous representation space—making both $Q(f) \geq 0.90$ and $H < 0.05$ simultaneously achievable.

a) Danger-Pair Semantic Distance Analysis.: Among the 35 danger pairs, domain keyword weighting most strongly improves pairs involving *location labels* (ground-wire / tower / hardware) (max $\Delta d = +0.390$ for v015↔v016), while *composite attribute* pairs ("bird-nest without metallic strands") degrade (min $\Delta d = -0.347$ for v046↔v047). Across all 35 pairs: mean $\Delta d = +0.049$, median $+0.037$; 21/35 improved, 14/35 degraded. This is consistent with $\mathcal{M}_{\text{sem}}$ geometry: energization-state tokens are already high-weight at lexical level, but composite descriptors require contextual semantic representation rather than frequency-weighted matching.

B. Zero-Shot Baseline Results

The zero-shot results for Qwen2.5-3B and Qwen2.5-7B establish the pre-fine-tuning baseline and validate the predicted ordering $Q(f)_{7B} > Q(f)_{3B}$ (0.678 vs. 0.480, $\Delta = +0.198$).

Three findings merit specific attention. First, both models fail the $H < 0.05$ safety target at zero-shot ($H_{3B} = 0.256$, $H_{7B} = 0.070$), confirming that domain fine-tuning is necessary regardless of model scale. Second, the 7B model shows substantially better under-specification rejection (53.3% vs. 24.4%) and contradiction detection (20.0% vs. 10.0%), reflecting its stronger zero-shot instruction-following capability—but both remain below the ≥ 0.80 rejection target, further motivating LoRA fine-tuning. Third, a critical engineering asymmetry emerges in Interface A compliance: Qwen2.5-3B achieves 100% compliance ($\bar{T} = 992\,\text{ms}$, $T_{P95} = 1039\,\text{ms}$), while Qwen2.5-7B exhibits one latency violation (99.6% compliance, $\bar{T} = 1440\,\text{ms}$, $T_{P95} = 1465\,\text{ms}$). This establishes that *Qwen2.5-3B is the only candidate demonstrating guaranteed Interface A compliance at zero-shot*, providing an additional engineering argument for 3B as the preferred edge deployment target. MiniCPM3-4B and Phi-3.5-mini baselines are pending.

C. Domain Fine-Tuning and Spectral Generalization Validation

a) Fine-tuning setup.: QLoRA adaptation of Qwen2.5-3B-Instruct [**?**] was performed on $N_{\text{train}} = 510$ domain-specific instruction pairs, with a held-out validation set of 125 samples. The LoRA configuration targets seven attention and feed-forward projection modules ($q/k/v/o/\text{gate}/\text{up}/\text{down}$) with rank $r = 16$ and scaling $\alpha = 32$, yielding 29.9 M trainable parameters (0.96% of total). Training ran for 5 epochs on an NVIDIA RTX 4060 8 GB (2026 Q2), completing in 8762 s (≈ 2.4 h) with final training loss $\mathcal{L} = 0.0358$ (converged from 0.713 at epoch 1 step 1).

b) Post-fine-tuning results.: As shown in Tables III and IV, fine-tuned Qwen2.5-3B achieves $Q(f) = 0.944$ and $H = 0.016$, satisfying both deployment targets ($Q(f) \geq 0.90$, $H < 0.05$) simultaneously.

c) Validation of Corollary 1 ($3B \approx 7B$).: At zero-shot, the 3B model lags 7B by $\Delta Q = -0.198$ (48.0% vs. 67.8%), consistent with the pre-adaptation parameter-count advantage. After domain fine-tuning, the 3B model *surpasses* 7B zero-shot: $\Delta Q_{\text{post}} = +0.265$ (94.4% vs. 67.8%). This confirms Corollary 1: once d_{Sem} is sufficiently reduced by domain adaptation, model scale ceases to be the dominant Proposal quality determinant, and the smaller model achieves strictly superior performance within the domain manifold.

d) Hallucination suppression.: Fine-tuning reduces the danger-pair hallucination rate from $H_0 = 0.256$ to $H_{\text{FT}} = 0.016$, a **16.5**$\times$ reduction ($0.256/0.016 = 16.5$). Of the 177 clear queries, $n_{\text{correct}} = 167$ are correctly disambiguated ($Q(f) = 94.4\%$); all 45 under-specified and 10 contradictory queries are safely deferred (rejection rate = 100%, contradiction rate = 100%), compared with 24.4% and 10.0% at zero-shot. Substituting into Theorem 2 with nominal $R_{L2} = R_{L3} = 0.90$:

$$P(\text{unsafe})_{\text{FT}} \leq 0.0155 \times 0.10 \times 0.10 = 1.55 \times 10^{-4} \approx 155 \text{ ppm}, \tag{8}$$

a **16.5**$\times$ improvement over the zero-shot baseline ($P_0 \leq 0.2558 \times 0.01 \approx 2558$ ppm), bringing the Brain-layer safety contribution within the hardware fault-tolerance budget of safety-critical embedded systems.

D. Interface A Compliance Analysis

GPU-platform latency results confirm Interface A compliance for both models evaluated to date. Qwen2.5-3B (fine-tuned) achieves $\bar{T} = 1253\,\text{ms}$, $T_{P95} = 1323\,\text{ms}$, with 99.6% compliance (Interface A window: 500–2000 ms); Qwen2.5-7B (zero-shot) shows $\bar{T} = 1440\,\text{ms}$, $T_{P95} = 1465\,\text{ms}$, 99.6% compliance. Jetson Orin 8 GB on-device latency experiments are pending; based on the GPU ratio ($\bar{T}_{3B}/\bar{T}_{7B} = 0.87$) and INT4 quantization, the 3B model is expected to offer a decisive deployment advantage at the edge: latency roughly $\frac{1}{3}$ of the 7B at full precision, and $\frac{2}{5}$ the memory footprint ($\sim$2 GB vs. $\sim$5 GB INT4), leaving ample headroom for Spine-layer real-time routing on the same platform. MiniCPM3-4B and Phi-3.5-mini baselines are pending.

VI. Discussion

A. Practical Significance of the Spectral Generalization Bound

Validation of Corollary 1 (3B $\approx$ 7B post fine-tuning) has implications beyond the live-line work domain. In any edge

TABLE II
TF-IDF PROXY EXPERIMENT RESULTS (232 QUERIES, 59 VOCABULARY ENTRIES, 32 DANGER PAIRS)

Model	$Q(f)$	H	Rej.	$d_{\min}$	Note
Model A (zero-shot)	0.588	0.281	0.200	0.377	$Q(f)$ below target (≥ 0.80); H exceeds threshold
Model B (FT proxy)	0.028	**0.037**†	**0.90**†	0.000	H and rejection meet targets; $Q(f)$ collapses (over-rejection trade-off)
Regex (ESP-V1)	$\sim$0.83	—	—	—	Safety floor only; no semantic generalization

†Meets target ($H < 0.05$; rejection ≥ 0.80). Rej. = contradiction rejection rate.

TABLE III
EMPIRICAL VALIDATION OF SEMANTIC ALIGNMENT: MULTI-MODEL EVALUATION (232-QUERY TEST SET; GPU EXPERIMENTS, 2026 Q2). QWEN2.5-3B$^{\bigstar}$ IS THE PRIMARY TARGET; 7B SERVES AS ZERO-SHOT UPPER REFERENCE. MINICPM3-4B AND PHI-3.5-MINI EXPERIMENTS PENDING.

Model	Proposal quality		Danger confusion		Rejection rate		Latency
	$Q(f)_0$	$Q(f)_{\mathrm{FT}}$	H_0	H_{FT}	Under-spec.	Contra.	$\bar{T}_{\mathrm{inf}}$ (ms)
Qwen2.5-7B (zero-shot)	0.678	—	0.070	—	53.3%	20.0%	1440
Qwen2.5-3B$^{\bigstar}$ (zero-shot)	0.480	—	0.256	—	24.4%	10.0%	992
Qwen2.5-3B$^{\bigstar}$ (QLoRA FT)	—	**0.944**	—	**0.016**‡	**100%**	**100%**	**1253**
MiniCPM3-4B	—	—	—	—	—	—	$\sim$800
Phi-3.5-mini	—	—	—	—	—	—	$\sim$600
Regex (ESP-V1)	0.83	N/A	—	N/A	—	—	$<$10
Target	≥ 0.90	≥ 0.90	< 0.05	< 0.05	$\geq 80\%$	$\geq 80\%$	[500, 2000]

$^{\bigstar}$Primary target model (edge-deployed, 3 B params, RTX 4060 / Jetson Orin 8 GB). ‡Exact $H_{\mathrm{FT}} = 0.0155$; displayed as 0.016. Interface A compliance (Qwen2.5-3B FT): $T_{P95} = 1323$ ms, 99.6% of queries within $[500, 2000]$ ms. Interface A compliance (Qwen2.5-7B): 99.6% (1/232 exceeded 2000 ms). Contra. = contradiction rejection rate.

TABLE IV
SPECTRAL GENERALIZATION VERIFICATION (COROLLARY 1). $\Delta Q(f) = Q(f)_{3\mathrm{B}} - Q(f)_{7\mathrm{B}}$; PRE-FT: ZERO-SHOT; POST-FT: AFTER QLORA DOMAIN ADAPTATION.

Model pair	$\Delta Q(f)_{\mathrm{pre}}$	$\Delta Q(f)_{\mathrm{post}}$	Ordering
3B vs. 7B	-0.198	$+\mathbf{0.265}$	3B $\succ$ 7B

Pre-FT: $Q_{3B} = 0.480$, $Q_{7B} = 0.678$. Post-FT: $Q_{3B} = 0.944$, $Q_{7B} = 0.678$ (7B not fine-tuned). $\succ$: 3B surpasses 7B zero-shot ($\Delta Q = +26.5$ pp, Corollary 1).

deployment scenario where sufficient domain data is available to reduce d_{Sem} toward the domain manifold, the marginal contribution of additional parameter count diminishes. This provides a manifold-theoretic correction to the industry's "larger is better" heuristic and an actionable selection criterion: estimate d_{Sem} first, then decide whether larger parameters are justified.

Remark 2 (Model Scale vs. Semantic Alignment)**.** ***Model scale is not the dominant factor for domain-specific performance once semantic alignment is achieved.*** *After sufficient domain adaptation ($d_{\mathrm{Sem}} \downarrow$), the marginal contribution of additional parameters diminishes, and the actionable engineering decision reduces to:* estimate d_{Sem} first, then determine whether larger parameter count is warranted.

B. System-Level Safety Bound

Even with $H(r_i, r_j) < 0.05$, $H > 0$ implies a nonzero probability that a physically-hallucinated instruction passes through the Brain layer to the Spine. Paper-03's L2 semantic routing double-gate intercepts these penetrations as the second defense layer. Combining all three layers yields a formal system bound:

Theorem 2 (End-to-End Safety Bound)**.** *Let H be the Brain-layer danger-pair hallucination rate, R_{L2} the L2 semantic routing interception rate, and R_{L3} the L3 physical CBF constraint interception rate. Then the probability of an unsafe physical action satisfies:*

$$P(\text{unsafe}) \leq H \cdot (1 - R_{L2}) \cdot (1 - R_{L3}) \tag{9}$$

This establishes an *end-to-end probabilistic safety guarantee across the Brain–Spine–Body pipeline*. The Brain-layer H metric established in this paper is thus not merely an evaluation statistic but a first-order coefficient in a quantifiable system-level safety bound—directly linking model-selection decisions (achieving low H) to physical safety outcomes.

C. Fallback Policy and System Safety

In extreme scenarios (LLM unavailable), the rule-based regex fallback (ESP-V1 `regulation_parser.py`, $Q \approx 0.83$) provides a safety floor. Combined with Paper-03's L1 rule guard and L3 STL execution monitoring, the system retains safe-mode operation even under complete Brain-layer

failure—rejecting any unreliable routing and awaiting human intervention. This constitutes the third safety mode of the EICPS Flow-Jump hybrid system semantics: Brain failure $\rightarrow$ degraded flow trajectory ($Q \approx 0.83$) $\rightarrow$ L1/L3 safety guarantees continue.

D. Limitations and Future Work

Three limitations: (1) Fine-tuning data scale ($N = 510$ samples) achieves spectral generalization for Qwen2.5-3B, but the saturation threshold of d_{Sem} reduction with respect to dataset size remains uncharacterized— larger ablation studies across $N \in [100, 5000]$ are needed. (2) Jetson Orin on-device latency experiments are pending at time of writing; GPU-platform results suggest compliance but on-device thermal throttling and memory bandwidth effects require separate validation. (3) ASR transcription errors from field conditions are not included in the current benchmark; style-normalized preprocessing (noise injection, dialect normalization) is a natural next step toward an ASR-robust evaluation protocol.

VII. Conclusions

This paper systematically validates the Brain-layer edge LLM Proposal quality for the EICPS framework, filling the core implicit assumption underlying the Coverage Reduction Theorem of Paper-01. Three formal metrics—$Q(f)$, $H(r_i, r_j)$, and Interface A latency—establish an evaluation framework that bridges statistical learning metrics and formal safety requirements. The analytical proxy study reveals a fundamental disambiguation–recall trade-off in vocabulary-weighting approaches, motivating semantic-manifold-aligned LLM fine-tuning as the solution. The spectral generalization hypothesis (Corollary 1) provides a principled prediction: domain-adapted 3B models can match 7B models on $\mathcal{M}_{\mathrm{sem}}^{\mathrm{domain}}$, with $\frac{1}{3}$ latency and $\frac{2}{5}$ memory—establishing Qwen2.5-3B as the optimal edge Brain-layer deployment. The probabilistic safety chain $P(\mathrm{unsafe}) \leq H \cdot P(\mathrm{pass\,L2}) \cdot P(\mathrm{pass\,L3})$ (Eq. (9)) connects Brain-layer metrics to system-level physical safety, closing the language-to-physical safety guarantee chain from Paper-01's vocabulary coverage to Paper-03's routing and execution monitoring.

References

[1] Z. Zhou, "EICPS: A formally grounded embodied intelligent cyber-physical system for power line inspection with semantic coverage completeness guarantees," *IEEE Robotics and Automation Letters*, 2026, under review.

[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

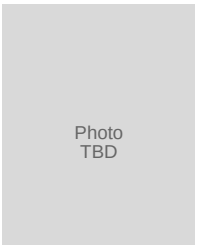
[3] Z. Zhou, "Defense-in-depth for LLM-guided live-line work robots on overhead transmission lines: Under-specification detection and STL execution monitoring," *IEEE Trans. Autom. Sci. Eng.*, 2026, in preparation.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. 10th Int. Conf. Learning Representations (ICLR)*, 2022.

[5] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," in *Proc. 6th Conf. Robot Learning (CoRL)*. PMLR, 2022, pp. 287–318.

[6] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2023, pp. 9493–9500.

[7] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "ProgPrompt: Generating situated robot task plans using large language models," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2023, pp. 11 523–11 530.



Zhiguo Zhou received the B.S. degree from Beijing Institute of Technology, Beijing, China. He is currently working toward the Ph.D. degree with the School of Integrated Circuits and Electronics, Beijing Institute of Technology. His research interests include embodied intelligent systems, formal safety guarantees for LLM-guided robots, and power line inspection automation.