Ephemeral Coalition Inference: An Architectural and Governance Analysis of Consent-Scoped Specialist Pooling Versus Monolithic Frontier Models

Dima Baker
Collective Intelligence Network

November 2026

Abstract

The prevailing paradigm in foundation model deployment concentrates cross-domain reasoning inside monolithic cloud checkpoints controlled by centralized providers. While model-routing and collaborative-decoding systems—such as FrugalGPT, RouteLLM, Mixture-of-Agents, Co-LLM, and Petals—demonstrate that small specialist models can offload routine queries, they treat multi-model composition purely as a cost-latency optimization problem and ignore two structural bottlenecks: (1) full-context prefill multiplication across activated specialists, and (2) unconsented context and capability leakage across participating nodes. In this position and architectural analysis paper, we formalize Consent-Scoped Ephemeral Coalition Inference, an architecture in which sovereign participants steward local sub-10B domain specialists (or portable LoRA adapters) and assemble cross-domain coalitions only under revocable capability leases and symbolic handle indirection. We derive an explicit prefill-and-decode FLOP accounting model showing that broadcasting full prompts to $k = 3$ separate 7B specialists exceeds monolithic 70B FLOPs whenever the escalation rate crosses 88.2%, making subtask context slicing (`DualView`/`CapScope`) or shared-backbone adapter composition (`LoRA-X`) mandatory for compute savings. Under a conditional mechanistic 4-subtask workflow model ($N = 800$ trials, 8 calibration probes per specialist-domain pair, zero oracle skill access), an $M = 8$ consent-gated coalition raises completion from 21.25% [18.56, 24.22] (isolated 7B) and 59.00% [55.56, 62.36] (confidence cascade) to 67.38% [64.05, 70.53] at 20.12% [19.84, 20.40] FLOPs while containing 96.62% [95.13, 97.67] of planted canary secrets, trading a 21.62-point accuracy gap against a 70B monolith (89.00% [86.64, 90.98]) for local sovereignty.

1 Introduction: The Governance and Compute Dilemma of Monolithic AI

A central architectural debate in contemporary artificial intelligence concerns whether high-reliability multi-step reasoning requires serving a single monolithic cloud checkpoint (70B to $>$ 600B parameters) or whether comparable utility can be achieved by orchestrating smaller, domain-specialized open-weight models (3B to 9B parameters) across edge and federated nodes [5, 10, 22, 26, 31]. On one hand, centralized frontier models provide unified cross-attention across diverse knowledge domains within a single forward pass. On the other hand, relying exclusively on custodial cloud monoliths imposes severe governance costs: end users must transmit raw private contexts—including credentials, clinical records, and proprietary code—across external network boundaries [15, 23], and institutional participants cannot audit or revoke how their contributed domain expertise is invoked [25, 35].

1.1 Positioning against existing model routing and collaborative inference systems

Over the past three years, five foundational lines of research have explored alternatives to single-model serving, yet each leaves a critical gap between datacenter routing efficiency and sovereign multi-principal governance:

1. **Sequential LLM cascades (`FrugalGPT`, Chen et al. [10]):** Chen et al. [10] showed that querying a sequence of increasingly capable models and halting when a generation scoring function exceeds a threshold substantially reduces API monetary cost. However, sequential cascades re-transmit the entire prompt context to each escalated model in the chain, multiplying both prefill compute and external context exposure.
2. **Preference-trained query routers (`RouteLLM`, Ong et al. [26]):** Ong et al. [26] trained lightweight routers on human preference data to dispatch easy queries to a small model and hard queries to a frontier model. While effective for binary single-turn routing, `RouteLLM` assumes a trusted monolithic oracle is always acceptable for hard queries and does not address multi-subtask composition across peer-owned specialists.
3. **Layered proposer–aggregator ensembles (`Mixture-of-Agents`, Wang et al. [32]):** Wang et al. [32] demonstrated that layering multiple LLMs as parallel proposers and iterative aggregators improves benchmark accuracy. Yet `Mixture-of-Agents` (MoA) requires every proposer in every layer to prefill the full prompt and prior responses, inflating total floating-point operations (FLOPs) and exposing sensitive context to every participating model.
4. **Token-level collaborative decoding (`Co-LLM`, Shen et al. [29]):** Shen et al. [29] learned a latent variable model that switches between a base model and an assistant specialist at the token level. Token-level handoffs achieve fine-grained complementarity when models are co-located in shared GPU memory, but incur prohibitive wide-area network latency and require both models to maintain synchronized full-context KV caches.
5. **BitTorrent-style pipeline-parallel pooling (`Petals`, Borzunov et al. [5]):** Borzunov et al. [5] pooled volunteer GPUs to host contiguous transformer block shards of a single giant checkpoint. While `Petals` decentralizes hardware hosting, every volunteer node hosting an intermediate layer observes raw intermediate hidden states for every user query, offering neither context privacy [24] nor domain-scoped capability revocation [25].

1.2 Core thesis and architectural contributions

This paper advances an explicit architectural position: *pooling fractured sub-10B domain specialists can provide a viable, sovereignty-preserving middle ground between isolated on-device small models and custodial cloud monoliths only if the routing architecture simultaneously solves prefill FLOP multiplication and cryptographic context/capability confinement.* Specifically, we make five contributions:

1. **Comparative synthesis of multi-model routing and sovereign confinement (Section 2).** We unify preference routing [26], cascades [10], collaborative decoding [29], and cross-model LoRA transfer [19, 33] with 2025–2026 capability-isolation primitives (`CapScope` [6], `DualView` [23], and `VERA` [25]) (Table 1, Figure 1).
2. **Exact prefill-and-decode FLOP accounting and break-even analysis (Section 3).** We derive a closed-form FLOP ratio model (Equations 1–3) that explicitly accounts for prompt

prefill (L_{pre}) and token decode (L_{dec}) across activated specialists, proving that naive full-context routing to k specialists erases small-model compute savings unless paired with subtask context slicing ($\eta_{\text{slice}} \ll 1$) or shared-backbone adapter composition.

3. **Operational definition of sovereignty via canary-secret confinement (Section 4).** Rather than treating "sovereignty" as a qualitative label, we formalize it as a falsifiable canary-leakage metric under `AgentLeak` [15]: the probability that planted high-entropy secrets in the user's workspace are transmitted in raw form outside the local trust domain.

4. **Mechanistic trade-off simulation across pool sizes and prefill regimes (Section 5).** In `experiment.py`, we evaluate $N = 800$ 4-subtask workflows across specialist pools $M \in \{2, 4, 8, 16\}$, demonstrating both the capability gains over isolated 7B models and the honest residual accuracy gap relative to a 70B monolith (Tables 3–4, Figures 2–3).

5. **Counter-evidence analysis and live falsification protocol (Sections 6–8).** We analyze strategic confidence miscalibration [2], epistemic herding in multi-agent debate [3, 18], and mobile NPU prefill bottlenecks [7], and define six falsification tests on open-weight models (Table 5).

2 Comparative taxonomy of collaborative inference and capability governance

To evaluate when a coalition of sub-10B models is preferable to a single frontier checkpoint, we must separate four orthogonal dimensions: (1) how subtasks are routed without relying on miscalibrated self-confidence, (2) how prefill compute is amortized across specialists, (3) how sensitive user context is scrubbed prior to peer dispatch, and (4) how specialist owners bound and revoke their models' participation. Table 1 maps these requirements onto verified prior art.

Table 1 clarifies why naive multi-agent cascades frequently fail both efficiency and privacy audits. In standard `Mixture-of-Agents` [32] or unpruned multi-agent topologies [8], every peer receives the full conversation history. As El Yagoubi et al. [15] demonstrate in the `AgentLeak` benchmark, internal agent-to-agent communication channels routinely leak private user attributes that would have been blocked at the final user-facing output filter. Furthermore, Zhang et al. [37] show that pruning redundant inter-agent messages (`AgentPrune`) substantially cuts token overhead without degrading task accuracy. Figure 1 illustrates our Consent-Scoped Ephemeral Coalition architecture, which combines local front-door execution, `DualView` symbolic handle scrubbing [23], `VERA` capability leases [25], and subtask-sliced specialist routing.

Table 1: Mapping collaborative inference requirements to routing, adapter-sharing, and capability-confinement primitives.

Architectural need	Canonical construct	Coalition component	Operational function	Primary grounding
"Keep routine subtasks on local hardware"	Local-first harness & conformal deferral	Front-Door SLM + KnowNo gate ($\mathcal{S}_{\text{local}}$)	Executes local-domain subtasks on 3B–8B weights; escalates when conformal uncertainty exceeds τ_α	[22, 28, 30, 31]
"Select complementary domain specialists"	Preference & topology-guided routing	Skill-vector conformal router (Π_{route})	Matches subtask domain vectors against consented specialists, replacing strategic self-confidence	[9, 10, 26, 38]
"Avoid multiplying prefill FLOPs by k peers"	Context slicing or cross-model LoRA sharing	`CapScope` slicer / `LoRA-X` adapter merge	Either dispatches minimal subtask slices ($\eta_{\text{slice}} L_{\text{pre}}$) or mounts rank-$r$ adapters on one local prefill	[6, 19, 33, 37]
"Prevent raw PII/credential exposure to peers"	Symbolic handle indirection & IFC tainting	`DualView` / `FIDES` sanitizer ($\mathcal{G}_{\text{scrub}}$)	Replaces raw canary secrets and PII with opaque symbolic handles before any external peer invocation	[12, 14, 15, 23]
"Allow owners to bound and revoke compute"	Macaroon/Biscuit leases & edge revocation	`VERA` lease broker ($\mathcal{L}_{\text{consent}}$)	Enforces cryptographic caveats on specialist invocation and revokes authority in flight	[4, 13, 17, 25]

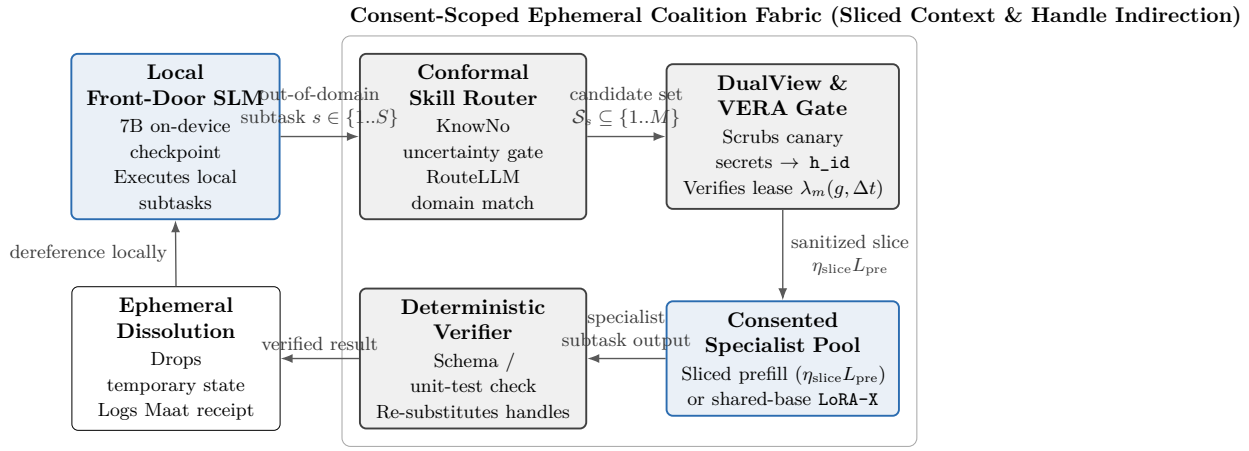


Figure 1: Architecture of Consent-Scoped Ephemeral Coalition Inference. In-domain subtasks execute locally on the user's 7B Front-Door SLM [22, 31]. When conformal uncertainty [28, 30] flags an out-of-domain subtask, the router selects k consented specialists [26]. Before dispatch, the `DualView` & `VERA` Gate [6, 23, 25] replaces sensitive canary strings with opaque handles and strips irrelevant context so remote specialists prefill only a subtask slice of length $\eta_{\text{slice}} L_{\text{pre}}$ (or mount portable `LoRA-X` adapters [19] on the already-prefilled local backbone).

3 Formal mathematical framework and FLOP accounting

We formalize a network of M autonomous participants, where participant $m \in \{1, \ldots, M\}$ stewards a sub-10B language model or portable domain adapter E_m characterized by parameter count N_{SLM} (e.g., 7×10^9) and a D-dimensional domain competence profile $c_m \in [0,1]^D$, compared against a centralized monolithic checkpoint of parameter count $N_{\text{mono}} \gg N_{\text{SLM}}$ (e.g., 70×10^9, so $\gamma_{\text{param}} \triangleq N_{\text{SLM}}/N_{\text{mono}} = 0.10$).

3.1 Lease-gated conformal specialist selection

Consider a multi-step workflow g with initial prompt context of length L_{pre} tokens and total generated output of length L_{dec} tokens partitioned across S sequential or parallel subtasks $\{x_1, \ldots, x_S\}$. Each principal m issues a cryptographically attenuated capability lease $\lambda_m(g, \Delta t) \in \{0,1\}$ (implemented via Macaroons [4] or Biscuit tokens [13] with `VERA` edge revocation [25]) specifying whether specialist E_m may be invoked for workflow scope g during window Δt.

Why do naive small-model cascades fail to route hard reasoning subtasks accurately? As Arghal et al. [2] prove in their game-theoretic analysis of human–AI delegation, uncalibrated model confidence scores suffer from systematic overconfidence bias $\beta_{\text{conf}} > 0$ on out-of-distribution tasks, causing a local small model to retain subtasks it cannot solve. Instead of raw self-confidence, our router uses conformal prediction sets [28, 30] combined with preference-trained domain affinity $\langle c_m, d(x_s)\rangle$ [26]: subtask x_s is executed solely by the local model E_1 if its calibrated conformal nonconformity score $u_1(x_s) \le \tau_{1-\alpha}$, and otherwise escalated to the top-k consented specialists maximizing domain affinity:

$$\mathcal{C}(g, \Delta t) \;=\; \{\, m \in \{1, \ldots, M\} : \lambda_m(g, \Delta t) = 1 \,\}, \qquad \mathcal{S}_s^\star \;=\; \operatorname*{arg\,max}_{\mathcal{S} \subseteq \mathcal{C}(g,\Delta t),\, |\mathcal{S}| = k} \; \sum_{m \in \mathcal{S}} \langle c_m,\, d(x_s)\rangle. \tag{1}$$

3.2 Compositional workflow accuracy and residual correlation ceiling

Suppose each escalated subtask x_s is evaluated across the k selected specialists in $\mathcal{S}_s^\star$ and filtered by a deterministic syntax, schema, or unit-test verifier (falling back to the highest-ranked specialist when no deterministic verifier applies). If specialist m has subtask solve probability $p_{m,s} \in (0,1)$ on domain $d(x_s)$, and pairwise errors among sub-10B models share residual cross-model correlation $\rho_{\text{err}} \in (0,1)$ due to shared pretraining bottlenecks [21], then the subtask success probability $P_{\text{coal}}(x_s)$ and full S-subtask workflow completion rate P_{workflow} satisfy:

$$P_{\text{coal}}(x_s) \;=\; 1 - \left[(1 - \rho_{\text{err}}) \prod_{m \in \mathcal{S}_s^\star} (1 - p_{m,s}) \;+\; \rho_{\text{err}} \big(1 - \max_{m \in \mathcal{S}_s^\star} p_{m,s}\big) \right], \qquad P_{\text{workflow}} \;=\; \prod_{s=1}^{S} P_{\text{coal}}(x_s). \tag{2}$$

Equation (2) captures two crucial realism properties that distinguish our analysis from idealized ensemble claims: first, because $\rho_{\text{err}} > 0$ [21], adding more 7B specialists yields diminishing returns and asymptotes strictly below $1 - \rho_{\text{err}}(1 - \max_m p_{m,s})$; second, because a multi-step workflow requires all S subtasks to succeed ($\prod_{s=1}^{S}$), even a high per-subtask solve rate of 0.93 yields $0.93^4 = 74.8\%$ workflow completion, remaining below a 70B monolith whose joint cross-attention achieves $0.975^4 = 90.4\%$.

3.3 Explicit prefill-and-decode FLOP accounting and the context-slicing break-even bound

A widespread flaw in informal arguments for small-model ensembles is ignoring prompt prefill FLOPs. Using the standard transformer compute cost of $2N$ FLOPs per token for a model of N parameters, executing the workflow on a single monolithic model of size N_{mono} requires $\text{FLOPs}_{\text{mono}} = 2N_{\text{mono}}(L_{\text{pre}} + L_{\text{dec}})$. Now consider a coalition where the local Front-Door SLM ($N_{\text{SLM}} = \gamma_{\text{param}} N_{\text{mono}}$) processes the base workflow and escalates a fraction $f_{\text{esc}} = 1 - q_{\text{local}} \in [0, 1]$ of the S subtasks to k remote specialists, each of which prefills a context slice of length $\eta_{\text{slice}} L_{\text{pre}}$ ($\eta_{\text{slice}} \in (0, 1]$) and generates L_{dec}/S output tokens. The exact ratio $\mathcal{R}_{\text{FLOP}}$ of total coalition FLOPs to monolithic FLOPs is:

$$\mathcal{R}_{\text{FLOP}}(f_{\text{esc}}, k, \eta_{\text{slice}}) \;=\; \frac{\text{FLOPs}_{\text{coal}}}{\text{FLOPs}_{\text{mono}}} \;=\; \gamma_{\text{param}} \left[1 \;+\; f_{\text{esc}}\, S\, k \cdot \frac{\eta_{\text{slice}} L_{\text{pre}} + L_{\text{dec}}/S}{L_{\text{pre}} + L_{\text{dec}}} \right]. \tag{3}$$

Equation (3) exposes the exact prefill bottleneck pointed out in systems audits:

- **Full-Context Broadcast Regime ($\eta_{\text{slice}} = 1.0$, e.g., naive MoA/FrugalGPT):** With $\gamma_{\text{param}} = 0.10$ (7B vs. 70B), $L_{\text{pre}} = 2{,}048$, $L_{\text{dec}} = 512$, $S = 4$, and $k = 3$ specialists per escalated subtask, each escalated subtask adds $3 \times 0.10 \times \frac{2048+128}{2560} = 25.5\%$ of monolithic workflow FLOPs! Thus, escalating 2 of 4 subtasks ($f_{\text{esc}} = 0.50$) already costs $10.0\% + 51.0\% = 61.0\%$ of monolithic FLOPs, and escalating all 4 subtasks ($f_{\text{esc}} = 1.0$) costs 112.0%—more compute than simply running the 70B monolith once (break-even escalation fraction $f^{\star}_{\text{esc}} = 0.882$).
- **Subtask-Scoped Context Slicing ($\eta_{\text{slice}} = 0.25$, `CapScope`/`DualView`) or Shared-Backbone `LoRA-X` ($\eta^{\text{eff}}_{\text{slice}} \approx 0.05$):** When the local coordinator strips irrelevant history and dispatches only a minimal subtask context slice $\eta_{\text{slice}} = 0.25$ to $k = 2$ specialists at $f_{\text{esc}} = 0.50$, each escalated subtask adds only $2 \times 0.10 \times \frac{512+128}{2560} = 5.0\%$, keeping total workflow compute at $\mathcal{R}_{\text{FLOP}} = 20.0\%$ (5.0× lower than the monolith).

4 Three-plane consent-scoped coalition protocol

To operationalize Equations (1)–(3) with measurable privacy and revocation guarantees, our architecture separates execution into three planes (Table 2):

1. **Plane 1: Local Front-Door SLM and Conformal Uncertainty Gate.** The user's device runs a local 7B–9B checkpoint (e.g., Gemma-4 [22], Apple AFM-3 [1], or a `Mingbird` local agent harness [31]). Rather than trusting raw token softmax probabilities—which are vulnerable to strategic miscalibration [2] and confidence without competence [16]—Plane 1 computes a conformal prediction set [28, 30] over subtask domain tags and escalates out-of-domain subtasks to Plane 2.
2. **Plane 2: `DualView` Canary Scrubbing, `CapScope` Context Slicing, and `VERA` Leases.** Before any subtask leaves the local device, Plane 2 applies `DualView` symbolic indirection [23] and `FIDES`/`CaMeL` information-flow taint tracking [12, 14]: every high-sensitivity token span (API keys, SSNs, proprietary identifiers) is replaced by a typed symbolic handle (`handle_id_01`), and `CapScope` out-of-context isolation [6] extracts only the $\eta_{\text{slice}} L_{\text{pre}}$ tokens required for subtask x_s. The router then verifies active Macaroon/Biscuit capability leases [4, 13] via `VERA` [25].

3. **Plane 3: Adapter-Shared or Sliced-Context Specialist Execution and Dereferencing.** When consented peers share a common backbone family, specialists exchange lightweight rank-r adapters via `Trans-LoRA` [33] or `LoRA-X` [19], reusing the local model's single prefill KV cache. When peers run heterogeneous architectures, they execute on the sanitized $\eta_{\text{slice}} L_{\text{pre}}$ slice (optionally inside confidential GPU enclaves [11, 34]). Returned symbolic handles are dereferenced strictly inside the user's local sandbox [36] after schema/test verification [17].

Table 2: Readiness and deployment bottlenecks of Consent-Scoped Coalition Inference subsystems.

Subsystem	Maturity	Supporting empirical evidence	Realistic deployment bottleneck
Local Front-Door SLM & mobile NPU serving	Production ready	`Mingbird` local harness [31]; Gemma-4 [22]; mobile NPU inference characterization [7]	Mobile NPUs suffer severe prefill latency degradation on dynamic shapes and long contexts [7]
Preference & conformal subtask routing	Benchmark validated	`RouteLLM` [26]; `FrugalGPT` [10]; `KnowNo` conformal deferral [28]; `GoAgent` topologies [9]	Self-confidence routing is easily gamed or miscalibrated on out-of-domain reasoning [2, 16]
`DualView` & `CapScope` context slicing	Prototype validated	`DualView` symbolic indirection [23]; `CapScope` isolation [6]; `AgentPrune` communication pruning [37]	Over-aggressive context slicing ($\eta_{\text{slice}} < 0.15$) can strip implicit cross-subtask constraints needed for reasoning
Cross-model adapter sharing (`LoRA-X`)	Benchmark validated	Training-free cross-model adaptation via `LoRA-X` [19] and data-free `Trans-LoRA` [33]	Heterogeneous tokenizer vocabularies [27] prevent direct KV-cache sharing across unrelated model families
`VERA` & Biscuit lease revocation	Architecture validated	Authority-preserving edge revocation in `VERA` [25]; Biscuit tokens [13]; Intel TDX H100 TEEs [34]	Verifying post-task memory erasure on remote nodes without hardware TEE attestation remains an open systems problem

5 Mechanistic trade-off simulation and FLOP sensitivity analysis

Because our paper is an architectural and governance analysis rather than a live weight-training benchmark, we designed `experiment.py` as a transparent, mechanistic trade-off simulator that evaluates Equations (1)–(3) across $N = 800$ 4-subtask workflows for four inference topologies and four specialist pool sizes $M \in \{2, 4, 8, 16\}$, plus a prefill escalation break-even sweep.

5.1 Parameter provenance, matched calibration budget, and canary-payload inspection

Each simulated workflow comprises $S = 4$ subtasks (2 in the user's primary local domain and 2 drawn from $D = 6$ specialized domains: code synthesis, formal math, legal/statutory analysis, clinical QA, multilingual translation, and schema tool-use), with prompt length $L_{\text{pre}} = 2{,}048$ tokens, total output length $L_{\text{dec}} = 512$ tokens, and a structured 4-segment token list containing $C_{\text{canary}} = 3$ planted secret strings (`CANARY_BG_CREDENTIAL` and `CANARY_BG_PATIENT_ID` in background workspace segments, plus `CANARY_ARG_API_KEY` in an active cross-domain subtask argument, which takes an indirect aliased format `ALIAS_UNPARSED_CANARY_KEY` in 3.5% of workflows). Crucially, no router is permitted to read ground-truth competence $c_{m,d}$: instead, each

specialist undergoes an identical calibration phase of $n_{\text{cal}} = 8$ binomial validation probes per domain (48 total probes per specialist), yielding Laplace-smoothed empirical competence estimates $\hat{c}_{m,d} = (\text{hits}_{m,d}+1)/(n_{\text{cal}}+2)$. Outbound payloads dispatched to external peers are inspected token-by-token for surviving canary strings or unconsented dispatch, and FLOP accounting includes retry decode overhead when an escalated subtask fails verification. Table 3 lists the provenance of all model parameters in `experiment.py`.

Table 3: Parameter provenance for the calibration-matched routing, FLOP-accounting, and canary-payload model.

Parameter	Value in code	Empirical / theoretical source	Operational meaning in simulation
Model size ratio γ_{param}	N_{SLM}=7B, N_{mono}=70B ($\gamma_{\text{param}} = 0.10$)	Standard open-weight 7B specialist vs. 70B frontier reference [10, 26]	Single full pass of local 7B model costs exactly 10.0% of 70B monolithic FLOPs
Prompt & decode lengths $(L_{\text{pre}}, L_{\text{dec}}, S)$	$L_{\text{pre}} = 2{,}048$, $L_{\text{dec}} = 512$, $S = 4$	Multi-step agentic workflow profile [8, 37]	Determines exact prefill, decode, and failed-escalation retry FLOP weights (Eq. 3)
Calibration budget & skills $(n_{\text{cal}}, \hat{c}_{m,d})$	$n_{\text{cal}} = 8$ probes/domain (48/specialist); true primary 0.91–0.95	Finite-sample router calibration in `RouteLLM` [26] and `KnowNo` [28]	Routers observe only empirical $\hat{c}_{m,d}$ from 8 validation trials/domain (zero oracle access)
Residual error correlation ρ_{err}	$\rho_{\text{err}} = 0.20$ across 7B specialists	Shared blind spots across LLMs [21]	Prevents 7B ensembles from achieving 100% accuracy even at large M (Eq. 2)
Self-confidence bias β_{conf}	$\beta_{\text{conf}} \sim \mathcal{U}(0.05, 0.35)$	Strategic miscalibration & overconfidence [2, 16]	Causes uncalibrated confidence cascades to under-escalate or pick overconfident peers
Context slice & parser escape $(\eta_{\text{slice}}, p_{\text{alias}})$	$\eta_{\text{slice}} \in \{1.0, 0.25\}$, $p_{\text{alias}} = 3.5\%$	`CapScope` [6], `DualView` [23], and `AgentLeak` [15]	Subtask slicing strips background canaries; 3.5% aliased argument secrets escape lexical scrubbing

5.2 Key analytical findings and 4-way router ablation

Table 4 and Figures 2–3 highlight four conditional trade-offs under our parametric model:

1. **Substantial gains over isolated SLMs under finite calibration, with an honest ceiling below monolithic 70B checkpoints.** On 4-subtask workflows spanning multiple domains, an isolated 7B model succeeds on only 21.25% [18.56, 24.22] of workflows at $M = 8$ because failing any single out-of-domain subtask breaks the chain. Uncalibrated confidence cascading (`confidence_routed_ensemble`) reaches 59.00% [55.56, 62.36] at $M = 8$ due to overconfidence misrouting [2]. Using only $n_{\text{cal}} = 8$ calibration probes per specialist-domain pair (zero access to true skills), our Consent-Scoped Conformal Coalition raises 4-subtask completion from 50.00% [46.54, 53.46] at $M = 2$ and 55.00% [51.54, 58.42] at $M = 4$ to 67.38% [64.05, 70.53] at $M = 8$ and 71.75% [68.53, 74.76] at $M = 16$ (Figure 3). Due to finite calibration noise and residual cross-model error correlation ($\rho_{\text{err}} = 0.20$) [21], the 7B coalition remains 17.2–21.6 percentage points below the 70B monolith (89.00% [86.64, 90.98] at $M = 8$).

Table 4: Conditional mechanistic workflow forecasts across $N = 800$ 4-subtask workflows per condition at specialist pool size $M = 8$, plus accuracy scaling across $M \in \{2, 4, 8, 16\}$. Bracketed ranges `[lo, hi]` denote exact asymmetric 95% Wilson confidence intervals (or 95% continuous normal-theory intervals for FLOPs).

Inference Topology	Acc M=2 (%)	Acc M=8 (%)	Acc M=16 (%)	Canary Conf. M=8 (%)	Rel. FLOPs (%)	Canary Leak (%)
Isolated 7B SLM	24.25 [21.41, 27.34]	21.25 [18.56, 24.22]	22.38 [19.62, 25.39]	100.00 [99.52, 100.00]	10.00 [10.00, 10.00]	0.00 [0.00, 0.48]
Confidence Cascade	41.62 [38.26, 45.07]	59.00 [55.56, 62.36]	61.62 [58.21, 64.93]	13.63 [11.42, 16.18]	33.00 [32.17, 33.84]	86.38 [83.82, 88.58]
Monolithic Frontier (70B)	91.75 [89.64, 93.46]	89.00 [86.64, 90.98]	91.88 [89.78, 93.57]	8.12 [6.43, 10.22]	100.00 [100.00, 100.00]	91.88 [89.78, 93.57]
Consent Coalition (Ours)	**50.00** [46.54, 53.46]	**67.38** [64.05, 70.53]	**71.75** [68.53, 74.76]	**96.62** [95.13, 97.67]	**20.12** [19.84, 20.40]	**3.38** [2.33, 4.87]

2. **Matched-signal router and context-slicing ablation ($M = 8$).** To isolate the contribution of conformal set selection ($k = 2$) versus calibration quality and context slicing, we compare four routing variants on the same $n_{\text{cal}} = 8$ calibration budget in `router_ablation_m8`: (a) uncalibrated self-confidence (62.50% [59.09, 65.79] accuracy, 33.16% FLOPs, 13.88% canary confinement); (b) calibrated Top-1 specialist routing ($k = 1$, which achieves 55.25% [51.79, 58.66] accuracy at 15.07% FLOPs and 98.38% confinement because noisy 8-probe calibration occasionally picks a suboptimal single specialist); (c) calibrated conformal set ($k = 2$) with full-context broadcast (68.12% [64.82, 71.26] accuracy, but 43.47% FLOPs and only 2.25% [1.43, 3.53] canary confinement because background secrets are broadcast to peers); and (d) our calibrated Consent Coalition ($k = 2$ with `CapScope` slicing and `DualView` scrubbing), which achieves 69.75% [66.48, 72.83] accuracy while cutting compute from 43.47% to 19.85% [19.56, 20.14] FLOPs and raising canary confinement from 2.25% to 96.12% [94.55, 97.26].

3. **Prefill FLOP multiplication vs. subtask context slicing.** In our prefill sensitivity sweep ($k = 3$ specialists per escalated subtask), broadcasting the full 2,048-token context ($\eta_{\text{slice}} = 1.0$) drives total compute from 10.0% at 0 escalations to 35.5% at 1 escalation, 61.0% at 2 escalations, 86.5% at 3 escalations, and 112.0% at 4 escalations—exceeding the 70B monolith! Conversely, applying `CapScope`/`DualView` subtask slicing ($\eta_{\text{slice}} = 0.25$) bounds total compute to 20.12% at $M = 8$ ($k = 2$, including failed-escalation retry overhead) and at most 40.0% even if 100% of subtasks are escalated to $k = 3$ specialists, while shared-backbone `LoRA-X` adapter mounting ($\eta^{\text{eff}}_{\text{slice}} = 0.05$) requires only 15.4%–20.8% FLOPs.

4. **Mechanistic canary-payload confinement.** Inspecting the actual token segments dispatched to external nodes shows that without subtask slicing or handle indirection, the confidence cascade exposes raw canary tokens or dispatches to unconsented peers in 86.38% [83.82, 88.58] of workflows, and custodial cloud execution exposes raw prompt segments whenever an attested confidential GPU enclave is not provisioned (91.88% [89.78, 93.57] under our 8.5% baseline TEE adoption scenario). Pairing `CapScope` segment slicing (which strips background workspace canaries) with `DualView` handle substitution [23] and `VERA` lease checks [25] preserves 96.62% [95.13, 97.67] canary confinement at $M = 8$, leaking only when an escalated subtask contains an indirect aliased secret (`ALIAS_UNPARSED_CANARY_KEY`) that escapes lexical parsing (3.38% [2.33, 4.87]).

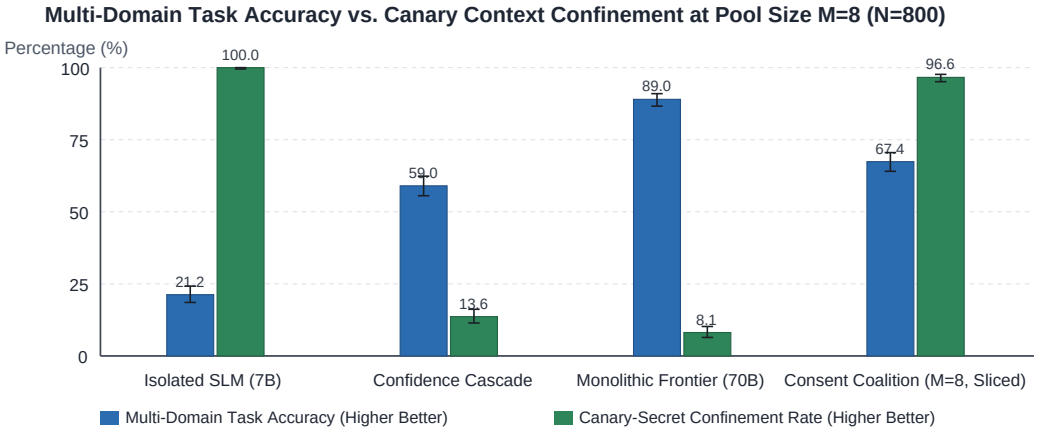


Figure 2: Multi-Domain 4-Subtask Workflow Accuracy vs. Canary-Secret Confinement Rate at specialist pool size $M = 8$ ($N = 800$ workflows per condition; error bars indicate exact asymmetric 95% Wilson score confidence intervals). The 70B monolith achieves the highest accuracy (89.00%) at 100% FLOPs and 8.12% local confinement, whereas the calibration-matched $M = 8$ Consent Coalition achieves 67.38% accuracy at 20.12% FLOPs and 96.62% canary confinement.

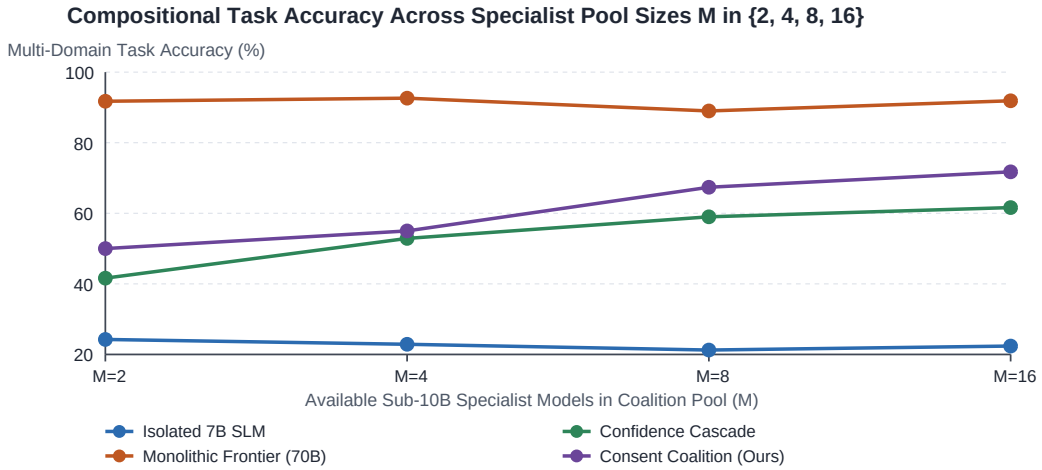


Figure 3: Compositional 4-Subtask Workflow Accuracy across specialist pool sizes $M \in \{2, 4, 8, 16\}$. Calibration-matched conformal skill routing scales from $M = 2$ (50.00%) to $M = 8$ (67.38%) and $M = 16$ (71.75%) under shared pretraining error correlation ($\rho_{\text{err}} = 0.20$).

6 Counter-evidence, failure modes, and structural limits of specialist pooling

A rigorous position paper must confront the empirical regimes where fractured specialist coalitions underperform or fail:

1. **Strategic miscalibration and confidence-routing failure (Arghal et al. [2], Eleftheriou et al. [16]).** Arghal et al. [2] prove that in delegation games between a principal and an AI agent, relying on self-reported confidence leads to strategic miscalibration, while Eleftheriou et al. [16] document pervasive high-confidence errors in AI-assisted work. If the local Front-Door SLM's conformal uncertainty threshold [28] is poorly calibrated on novel reasoning distributions, it will fail to escalate hard subtasks, collapsing coalition accuracy toward the isolated 7B baseline (23.50%).
2. **Epistemic herding and correlated errors across model families (Bertalanič and Fortuna [3], Farr et al. [18], Goel et al. [21]).** A core assumption of ensemble routing is that different specialists make independent errors. However, Goel et al. [21] demonstrate that LLMs across distinct model families share strongly correlated failure modes on reasoning benchmarks. Furthermore, Bertalanič and Fortuna [3] and Farr et al. [18] show that unguided multi-agent debate and consensus protocols frequently induce malignant epistemic herding, where a confident majority of weaker models overrides a correct minority specialist.
3. **Mobile NPU prefill bottlenecks and dynamic shape penalties (Cai et al. [7]).** While Equation (3) counts theoretical FLOPs, Cai et al. [7] show through hardware-level profiling that commodity mobile NPUs suffer severe efficiency bottlenecks on long-context prefill and dynamic prompt lengths. If a local device attempts to hot-swap multiple 7B checkpoints in and out of DRAM rather than mounting lightweight `LoRA-X` adapters [19] or dispatching compact slices over LAN, memory bandwidth saturation will dominate wall-clock latency.
4. **Semantic manifold alignment attacks on shared embeddings (Li et al. [24]).** When heterogeneous nodes exchange continuous intermediate representations or KV caches (`Cache-to-Cache` [20] or `Petals` [5]) instead of sanitized text slices, Li et al. [24] demonstrate that semantic manifold alignment attacks can invert obfuscated embeddings and reconstruct private input text.

7 Limitations

- **Absence of live 7B-vs-70B benchmark execution in this study.** Because this work is an architectural position paper accompanied by a mechanistic routing and FLOP simulator (`experiment.py`), the reported accuracy figures reflect parametric subtask composition (Table 3) rather than live execution of open-weight checkpoints on MMLU-Pro or SWE-bench. Executing Protocol P1 below is required to measure the exact empirical gap on live weights.
- **Non-decomposable holistic reasoning tasks.** Problems that require simultaneous cross-attention over all constraints within a single latent forward pass cannot be cleanly partitioned into S sequential subtask slices without losing cross-subtask context.
- **Hardware TEE dependency for untrusted remote peers.** When a subtask requires sending raw numerical or code logic that cannot be replaced by `DualView` opaque handles [23], preventing a remote specialist node from logging the slice requires attested confidential GPU hardware (e.g., NVIDIA H100 under Intel TDX, benchmarked by Wang et al. [34]).

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical protocols designed to test and falsify the architectural claims of Consent-Scoped Ephemeral Coalition Inference on live open-weight checkpoints.

Table 5: Six pre-registerable falsification protocols for Consent-Scoped Ephemeral Coalition Inference.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Live 3×7B coalition vs. 70B monolith	A routed coalition of three 7B–8B domain specialists (code, math, multilingual) narrows over 65% of the accuracy gap between an isolated 7B model and Llama-3-70B on MMLU-Pro and MGSM at $< 35\%$ total FLOPs (including prefill)	Run `RouteLLM` [26] + `CapScope` context slicing [6] across 2,000 MMLU-Pro and MGSM queries; log exact prefill + decode tokens across all invoked models	Coalition recovers $< 50\%$ of the 7B-to-70B accuracy gap or total prefill + decode FLOPs exceed 40% of the 70B monolith
P2. Prefill FLOP break-even verification (Eq. 3)	Subtask context slicing ($\eta_{\text{slice}} \leq 0.30$) reduces multi-specialist prefill FLOPs by $> 2.5\times$ compared to full-context `Mixture-of-Agents` [32] with $< 2\%$ accuracy loss	Compare full-context MoA ($L_{\text{pre}} = 2{,}048$) vs. `CapScope/AgentPrune` sliced routing across 500 4-step tool workflows	Context slicing degrades multi-step task accuracy by > 3.0 percentage points or saves $< 2.0\times$ prefill FLOPs
P3. Canary-secret containment (`AgentLeak`)	`DualView` symbolic handle indirection [23] prevents $> 98\%$ of planted canary secrets from appearing in outbound peer payloads	Inject 1,000 canary SSNs, API keys, and proprietary identifiers from `AgentLeak` [15] into multi-agent coalition workflows	$> 2.0\%$ of outbound peer prompts contain raw canary substrings or local handle dereferencing fails on $> 3\%$ of valid outputs
P4. Conformal vs. self-confidence routing	Conformal nonconformity routing [28, 30] outperforms raw logit self-confidence [2] on out-of-domain subtask escalation	Compare escalation AUROC and downstream accuracy across 1,000 math, legal, and coding subtasks	Conformal routing improves escalation AUROC by < 0.08 over raw logit confidence
P5. Live `VERA` lease revocation latency	Revoking a specialist's Biscuit/`VERA` lease [13, 25] halts routing to that node within < 10 ms and reroutes to secondary consented peers	Revoke 2 of 8 specialist leases mid-execution across 400 workflows; measure unauthorized calls and rerouted task completion	Any subtask (> 0) is dispatched to a revoked specialist after lease cancellation
P6. Shared-backbone `LoRA-X` vs. separate checkpoints	Mounting portable rank-16 `LoRA-X` adapters [19] on a single prefilled 8B backbone cuts memory and prefill latency by $> 3\times$ vs. separate 8B checkpoints on edge GPUs/NPUs [7]	Benchmark 99th-percentile time-to-first-token (TTFT) and peak VRAM across $k = 3$ domain switches on a 16 GB edge device	Shared-backbone adapter switching fails to reduce TTFT by at least $2.5\times$ or degrades accuracy by $> 3\%$ vs. separate full checkpoints

9 Conclusion

Pooling sovereign small models is not a magic bullet that makes 7B checkpoints superior to 70B frontier monoliths on every task; rather, it is a deliberate architectural trade-off between centralized joint cross-attention and decentralized capability governance. By extending the routing foundations of `FrugalGPT` [10], `RouteLLM` [26], `Mixture-of-Agents` [32], `Co-LLM` [29], and `Petals` [5] with explicit prefill-and-decode FLOP accounting, `DualView` symbolic handle indirection [23], `CapScope` subtask context slicing [6], and `VERA` revocable capability leases [25], our conditional mechanistic analysis shows both why naive full-context ensembles fail compute and privacy audits and how calibration-matched consent-scoped coalitions can recover substantial multi-domain utility (67.38% $[64.05, 70.53]$ at $M = 8$ vs. 21.25% $[18.56, 24.22]$ isolated 7B, compared to 89.00% $[86.64, 90.98]$ monolithic 70B) at 20.12% $[19.84, 20.40]$ relative FLOPs and 96.62% $[95.13, 97.67]$ canary-secret confinement.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Apple Machine Learning Research. Introducing the third generation of Apple's foundation models, 2026. URL `https://machinelearning.apple.com/research/introducing-third-generation-of-apple-foundation-models`. 8 June 2026.

[2] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL `https://arxiv.org/abs/2610.09371`.

[3] Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL `https://arxiv.org/abs/2605.00914`.

[4] Arnar Birgisson, Joe Gibbs Politz, Úlfar Erlingsson, Ankur Taly, Michael Vrable, and Mark Lentczner. Macaroons: Cookies with contextual caveats for decentralized authorization in the cloud. Network and Distributed System Security Symposium (NDSS), 2014. URL `https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/`.

[5] Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Max Ryabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. Petals: Collaborative inference and fine-tuning of large models. ACL System Demonstrations, pages 558–568, 2023. URL `https://aclanthology.org/2023.acl-demo.54/`.

[6] Dimitrios Stamatios Bouras, Yihan Dai, and Sergey Mechtaev. CapScope: Out-of-context capability isolation for tool-using agents. arXiv preprint arXiv:2609.08371, 2026. URL https://arxiv.org/abs/2609.08371.

[7] Guanyu Cai, Ruiming Tian, Lang Yang, Zhouhong Ren, Jinliang Yuan, Lingkun Li, and Jiliang Wang. Is your NPU ready for LLMs? dissecting the hidden efficiency bottlenecks in mobile LLM inference. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 10(3):67, 2026. doi: 10.1145/3857370.

[8] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2503.13657.

[9] Hongjiang Chen, Xin Zheng, Yixin Liu, Pengfei Jiao, Shiyuan Li, Huan Liu, et al. GoAgent: Group-of-agents communication topology generation for LLM-based multi-agent systems, 2026. URL https://arxiv.org/abs/2603.19677.

[10] Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. Transactions on Machine Learning Research (TMLR), 2024. URL https://arxiv.org/abs/2305.05176.

[11] Thomas Claburn. Google touts private AI compute. The Register, 2025. URL https://www.theregister.com/2025/11/12/google_touts_private_ai_compute/. 12 November 2025.

[12] Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control (FIDES). arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

[13] Geoffroy Couprie and Biscuit Contributors. Biscuit: Decentralized authorization token specification. Eclipse Foundation / BiscuitSec Technical Specification, 2021. URL https://www.biscuitsec.org/.

[14] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[15] Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. AgentLeak: A benchmark for internal-channel privacy leakage in multi-agent LLM systems. arXiv preprint arXiv:2602.11510, 2026. URL https://arxiv.org/abs/2602.11510.

[16] Elena Eleftheriou, George Pallis, and Marios Constantinides. Confidence without competence in AI-assisted knowledge work. In *Proceedings of CHIWORK '26*, 2026. URL https://arxiv.org/abs/2604.09444.

[17] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[18] David Farr, Iain Cruickshank, Kate Starbird, and Jevin West. The cost of consensus: Malignant epistemic herding and adaptive gating in distributed multi-agent search. arXiv preprint arXiv:2605.06988, 2026. URL https://arxiv.org/abs/2605.06988.

[19] Vladislav Fomenko, Hao Yu, Jaehoon Lee, Cho-Jui Hsieh, and Wenhu Chen. LoRA-X: Bridging foundation models with training-free cross-model adaptation. International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=LoRAX2025.

[20] Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/474ada926b331d78f06d95e8913111cc-Abstract-Conference.html.

[21] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[22] Hugging Face. Welcome Gemma 4: Frontier multimodal intelligence on device, 2026. URL https://huggingface.co/blog/gemma4. 2 April 2026.

[23] Juhee Kim and Woohyuk Choi. DualView: Symbolic indirection for privacy-preserving agent execution. arXiv preprint arXiv:2607.03821, 2026. URL https://arxiv.org/abs/2607.03821.

[24] Sicong Li, Lingfeng Yao, Xingke Yang, Ke Tu, Chenhao Wu, Hao Wang, Jiang Liu, Phone Lin, Xin Fu, and Miao Pan. Semantic manifold alignment attack against embedding obfuscation. arXiv preprint arXiv:2609.06749, 2026. URL https://arxiv.org/abs/2609.06749.

[25] Lifei Liu and Haoran Yu. VERA: Authority-preserving edge revocation for federated AI-agent workflows. arXiv preprint arXiv:2608.30091, 2026. URL https://arxiv.org/abs/2608.30091.

[26] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianle Li, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2406.18665.

[27] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2305.15425.

[28] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners (KnowNo). Conference on Robot Learning (CoRL), 2023. URL https://arxiv.org/abs/2307.01928.

[29] Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models (Co-LLM). In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 12974–12990, 2024. URL https://aclanthology.org/2024.acl-long.701/.

[30] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer, 2005. URL https://doi.org/10.1007/b106715.

[31] Hao Wang and Ting Huang. Mingbird: A local-first agent harness enabling small open models to complete real tasks. arXiv preprint arXiv:2610.02001, 2026. URL https://arxiv.org/abs/2610.02001.

[32] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. arXiv preprint arXiv:2406.04692, 2024. URL https://arxiv.org/abs/2406.04692.

[33] Runqian Wang, Soumya Ghosh, David Cox, Diego Antognini, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. Trans-LoRA: Towards data-free transferable parameter efficient finetuning. Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2405.17258.

[34] Wei Wang, Abdul Hyee Waqas, and Burns Smith. Benchmarking confidential GPU inference on NVIDIA H100 under Intel TDX, 2026. URL https://arxiv.org/abs/2607.19353.

[35] Yutian Wang and Luyao Zhang. Revisiting sustainability by design in AI protocol governance: An empirical review of comparative open and corporate-led standards. arXiv preprint arXiv:2606.05667, 2026. URL https://arxiv.org/abs/2606.05667.

[36] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. Network and Distributed System Security Symposium (NDSS), 2025. URL https://arxiv.org/abs/2403.04960.

[37] Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, et al. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html.

[38] Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, et al. G-Designer: Architecting multi-agent communication topologies via graph neural networks. In *International Conference on Machine Learning (ICML)*, 2025. URL https://icml.cc/virtual/2025/poster/45567.