EuReason: A Federated Neuro-Symbolic Framework for Verifiable Reasoning over Sovereign Data

Anonymous ACL submission

Abstract

Large language models (LLMs) have recently shown strong improvements in multi-step reasoning through techniques such as chain-of-thought prompting and inference-time compute scaling. However, two limitations hinder their deployment in high-stakes domains: (1) reasoning traces produced by LLMs are not guaranteed to faithfully reflect the model's internal computation, and (2) many valuable knowledge sources reside in regulated or organizational data silos that cannot be centrally aggregated. These challenges motivate architectures that combine verifiable reasoning with privacy-preserving access to distributed knowledge. This paper analyzes recent advances across four research areas—sparse Mixture-of-Experts architectures, inference-time reasoning, neuro-symbolic verification, and federated learning—and identifies open challenges in integrating them into trustworthy AI systems. We propose **EuReason**, a conceptual framework for verifiable reasoning over distributed data. EuReason organizes the problem into three interacting layers: (i) a sovereign data layer enabling federated access to heterogeneous knowledge sources via virtual knowledge graphs, (ii) a federated learning fabric designed to handle non-IID data distributions across decentralized institutions, and (iii) a neuro-symbolic reasoning layer that combines LLM generation with argumentative and logical verification of intermediate reasoning steps. Through a structured synthesis of recent work, we identify eight research gaps spanning distributed expert routing, faithful reasoning traces, scalable symbolic verification, and knowledge grounding across federated sources. We argue that addressing these gaps is necessary for deploying reasoning systems that are verifiable, privacy-preserving, and auditable in domains such as healthcare, governance, and scientific decision support.

1 Introduction: Verifiable Reasoning over Distributed Knowledge

Recent advances in Large Language Model (LLM) reasoning—from chain-of-thought prompting (Wei et al., 2022) to inference-time compute scaling (OpenAI, 2024)—have demonstrated that extended internal deliberation can solve mathematical olympiad problems and multi-step verification tasks. This transition reflects the shift from "System 1" (fast, intuitive, probabilistic) to "System 2" (slow, deliberate, verifiable) processing (Kahneman, 2011).

However, deploying these capabilities in high-stakes domains—healthcare diagnostics, legal compliance, scientific decision support—exposes two limitations. First, generated reasoning traces are not guaranteed to faithfully reflect the model's internal computation, a problem known as **unfaithful explanation** (Adadi and Berrada, 2018). Second, many valuable knowledge sources reside in regulated or organizational data silos (hospitals, government archives, industrial databases) that cannot be centrally aggregated due to privacy regulations such as GDPR and the EU AI Act (European Parliament and Council of the European Union, 2024). These are not uniquely European concerns: any domain where data governance, auditability, and institutional boundaries matter faces the same technical challenges.

1.1 The Data Access Problem

The reliance on centralized data aggregation faces a **Data Wall**—not merely the exhaustion of public textual tokens (Villalobos et al., 2022), but the inaccessibility of high-value, domain-specific data locked within organizational silos due to legitimate regulatory frameworks. Training on synthetic data to compensate generates "model collapse," where variance and creativity degrade over generations (Shumailov et al., 2024).

As Hoos and the CLAIRE initiative have argued, collaborative, decentralized infrastructure is needed to support AI research under ethical and privacy constraints (Tornede et al., 2023). More broadly, the concept of **Sovereign AI**—a nation's or organization's ability to develop, deploy, and govern AI systems using its own infrastructure and data (de Ocáriz Borde, 2025)—captures the requirement that reasoning systems must operate *over* distributed data rather than *on* centralized copies of it.

1.2 The Verification Gap

Despite massive parameter counts (DeepSeek-V3: 671B (DeepSeek-AI, 2025); Llama 4: 400B (AI, 2025)), current systems exhibit "jagged intelligence" (Dell'Acqua et al., 2023): PhD-level performance on specialized examinations (AIME 2025: 86.7%) alongside failure on elementary reasoning tasks (ARC-AGI-2: 4%) (Chollet et al., 2024). This suggests that current architectures lack world grounding and causal understanding (Pearl and Mackenzie, 2018), motivating work on structured verification (Freedman et al., 2025) and decentralized knowledge integration (Xiao et al., 2019).

1.3 Contribution and Outline

Taken together, the data access problem and the verification gap define a single overarching challenge: *how can AI systems reason verifiably over knowledge that must remain distributed?* Addressing this challenge requires simultaneous progress on three fronts—efficient access to sovereign data (Sections 2 and 5), trustworthy verification of reasoning (Sections 3 and 4), and grounding in authoritative, distributed knowledge sources (Section 6). EuReason (Section 8) integrates these fronts into a dependency graph of eight research gaps whose resolution would enable verifiable, privacy-preserving reasoning for high-stakes domains.

This paper makes three contributions beyond literature synthesis.

(C1) Dependency analysis. We identify eight research gaps across distributed architecture, reasoning verification, and knowledge grounding, and show that they form a directed dependency structure: progress on trust infrastructure (G1) and ontology construction (G7) gates downstream advances in federated learning (G2, G6) and verification (G3–G5). This structure prioritizes the research agenda.

(C2) Technical grounding of federated MoE. We provide a structurally grounded argument—based on conditional computation, expert independence, and heterogeneous specialization—for why MoE architectures are suited to distributed sovereignty settings, distinguishing this from suggestive analogy (Section 2).

(C3) Evaluation framework. We propose concrete metrics for each gap (Table 3), specify inter-layer interfaces with failure modes, and compare EuReason against four neighboring paradigms (RAG, tool use, modular pipelines, conventional trustworthy AI) to clarify its scope and limitations (Section 9).

Nature of EuReason. EuReason is a *conceptual dependency graph*—not a taxonomy (which would merely classify), nor a systems blueprint (which would require implementation details we cannot yet provide). Its value lies in making explicit which problems must be solved before others can be meaningfully addressed, and in providing evaluation criteria (Table 3) against which future work can measure progress.

Scope. While we frame EuReason using European regulatory constraints (GDPR, EU AI Act) as concrete instantiations, the technical challenges—verifiable reasoning, federated knowledge access, non-IID learning—apply to any domain with institutional data boundaries. The framework is most immediately applicable to *trustworthy domain-specific systems* in healthcare, governance, and legal compliance; whether it extends to general intelligence remains an open question that we do not claim to resolve.

Throughout, we use explicit epistemic markers to distinguish *established results* (demonstrated empirically or formally), *extrapolations* (logical extensions of established results under new assumptions), and *conjectures* (plausible but unvalidated proposals).

The paper is organised so that each section addresses a specific subproblem of the overarching challenge. Section 2 asks *how models can scale efficiently without centralising data* (→ gaps G1–G2). Section 3 asks *how reasoning quality scales at inference time and why current traces cannot yet be trusted* (→ gaps G3–G4). Section 4 asks *how reasoning steps can be formally verified* through neuro-symbolic argumentation (→ gaps G3–G5). Section 5 asks *how models learn from data that cannot leave its jurisdiction* via federated learning (→ gaps G1–G2, G6). Section 6 asks *how claims can be grounded in authoritative sources without centralising them* via virtual knowledge graphs (→

gaps G6–G8). Section 7 broadens the scope to embodied, cultural, and sustainability dimensions. Section 8 synthesises these analyses into the **EuReason** dependency graph, specifying inter-layer interfaces, failure modes, and evaluation metrics. Section 9 positions EuReason against neighbouring paradigms and Section 10 distils priorities and impact.

1.4 Methodology

We decompose the problem into three dimensions: (i) architectural efficiency and data access, (ii) reasoning verification and trust, and (iii) distributed learning and sovereignty. For each, we survey recent advances (2023–2026) from peer-reviewed venues and preprints, assess what problems each approach solves, what assumptions it requires, and what limitations remain. A research gap is identified when no surveyed approach addresses a limitation without introducing assumptions incompatible with privacy-preserving, auditable deployment. We organize gaps into the EuReason dependency graph (Section 8).

2 The Architectural Transition: From Dense to Sovereign Sparse Systems

The first subproblem is architectural: current dense models scale poorly, and their reliance on centralised data is incompatible with sovereignty requirements. We analyse how sparse Mixture-of-Experts architectures mitigate the efficiency problem and derive the gaps that remain for distributed deployment.

2.1 The Limits of Dense Scaling and the "Data Wall"

The primary constraint is the "Data Wall"—the exhaustion of high-quality public training data (Villalobos et al., 2022). Training on synthetic data to compensate generates "model collapse" (Shumailov et al., 2024).

In response, the industry has shifted toward **sparse Mixture-of-Experts (MoE)** architectures, where a learned gating function routes inputs to specialized expert networks (Shazeer et al., 2017). For input x:

$$y = \sum_{k=1}^{n} g_k(x) \cdot E_k(x) \tag{1}$$

where $g_k(x) = \text{softmax}(W_g^T x)_k$ is the gating weight and $E_k(x)$ is the output of the k-th expert. This allows models to scale to trillions of parameters while only activating a fraction (e.g., 2%–4%) during inference (AI, 2025)(DeepSeek-AI, 2025). Table 1 provides a comparative analysis of contemporary sparse architectures, including our proposed Sovereign-MoE.

2.2 Sovereign MoE: A European Proposal

While MoE improves efficiency, standard implementations still reside in centralized data centers. We propose a **Sovereign Federated Mixture-of-Experts (FedMoE)** architecture. In this paradigm, "experts" are not just sub-networks within a GPU cluster but distinct, decentralized models residing in *sovereign data spaces* (e.g., hospitals, national archives, industrial servers).

The structural suitability of MoE for federated sovereignty extends beyond analogy. Three architectural properties make MoE compatible with distributed deployment: **(i) Conditional computation**: MoE already routes inputs to subsets of parameters, so replacing intra-cluster routing with cross-institution routing preserves the same mathematical interface (Equation 1) while changing only the communication topology. **(ii) Expert independence**: during inference, activated experts compute independently given the routed input—no cross-expert attention or shared hidden states are required within a single forward pass, enabling geographic distribution without architectural modification. **(iii) Heterogeneous specialization**: existing MoE training naturally produces experts that specialize on different data distributions (Shazeer et al., 2017); in a federated setting, this specialization aligns with the inherent non-IID structure of sovereign data (e.g., German medical records vs. Spanish manufacturing data).

Epistemic status: Properties (i)–(ii) are established results from MoE literature. Property (iii) is an extrapolation: while expert specialization is empirically observed in centralized MoE, whether this specialization remains stable under federated training with non-IID data is an open question that constitutes part of gap G2.

This aligns with **Sovereign AI** (de Ocáriz Borde, 2025)—the requirement that models be anchored in local cultural and legal contexts. By creating a federation where the routing layer is shared but experts are sovereign, a collective intelligence can respect the non-IID nature of multi-jurisdictional data. The Helmsman framework (Li et al., 2025) demonstrates that multi-agent collaboration can au-

Table 1: Sparse Architectural Paradigms (2024–2026)

Architecture	Total	Active	Context	Routing
Llama 4 Maverick	400B	17B	1M	Hybrid
DeepSeek-V3	671B	37B	128K	MH Latent
OLMoE-1B-7B	7B	1B	64K	Fine-grained
Sovereign-MoE*	Var.	Var.	Virtual	Fed. Virtual

* Proposed in this work (conceptual).

tonomously synthesize such FL systems, enabling distributed expert training without centralized coordination.

Gap Summary (Architectural Dimension): The transition to sparse MoE addresses efficiency but does not resolve: (G1) routing queries to geographically distributed experts without a central coordinator, and (G2) training sovereign experts on non-IID data without compromising local data governance.

3 Inference-Time Scaling: From System 1 to System 2

Even if architectural challenges (G1, G2) were resolved, trustworthy AI requires verifiable reasoning. We assess inference-time scaling and identify verification gaps.

3.1 Mechanisms of Deliberative Reasoning

The "System 2" transition is driven by **Inference-Time Compute Scaling**. Models like DeepSeek-R1 (DeepSeek-AI, 2025) allocate compute to explore reasoning paths, verify intermediate steps, and backtrack upon errors, trained via **Reinforcement Learning with Verifiable Rewards (RLVR)**. The scaling law for inference implies that for difficult problems, increasing test-time compute (sampling k paths) yields better returns than scaling model size (Brown et al., 2024):

$$\text{PassRate}(k) \approx 1 - (1 - P_{\text{base}})^k \quad (2)$$

Snell et al. (Snell et al., 2024) show that compute-optimal strategies—adaptively allocating test-time compute per prompt difficulty—improve efficiency by more than 4x. Zhu et al. (Zhu et al., 2025) formalize this as a bandit problem, achieving up to 11% accuracy improvements. The reasoning paradigm landscape spans Chain-of-Thought (Wei et al., 2022), Tree-of-Thought, Meta-CoT (Xiang et al., 2025), and RLVR-CoT approaches, each with distinct tradeoffs between computational overhead and verification capability (Table 2).

3.2 The Trust Deficit in Black-Box Reasoning

While effective, these reasoning traces remain opaque. There is no guarantee that generated reasoning matches internal state—a problem of **unfaithful explanation** (Adadi and Berrada, 2018). For critical applications, probabilistic "thoughts" are insufficient; we need *verified reasoning* (Colelough and Regli, 2025).

Gap Summary (Reasoning Dimension): Inference-time scaling improves performance but does not resolve: (G3) guaranteeing faithful reasoning traces, and (G4) verifying reasoning against external knowledge without centralizing it.

4 The Trust Layer: Neuro-Symbolic Argumentation and Verification

To address G3 (faithful reasoning) and G4 (knowledge verification), we analyze neuro-symbolic approaches that replace opaque "thoughts" with contestable "arguments."

4.1 Argumentative Large Language Models (ArgLLMs)

Building on Toni et al.'s Argumentative AI (Freedman et al., 2025), we propose integrating **ArgLLMs** into inference. When generating a reasoning step, the model also generates a **Quantitative Bipolar Argumentation Framework (QBAF)**—extending Dung's framework (Dung, 1995)—with arguments $\mathcal{A}$ and relations of attacks $\mathcal{R}^-$ and supports $\mathcal{R}^+$. A step s_t is accepted only if its dialectical strength exceeds threshold τ:

$$\sigma(s_t) = f\big(\tau(s_t), \text{Attacks}(s_t), \text{Supports}(s_t)\big) \quad (3)$$

where $\tau(s_t)$ is the intrinsic strength. The QBAF provides a formal, human-readable justification for *why* a decision was made, enabling **contestability** as required by Articles 13–14 of the EU AI Act (European Parliament and Council of the European Union, 2024).

Interaction with neural generation. Integration operates through a *generate-then-verify*

Table 2: Reasoning Paradigms (2024–2026)

Paradigm	Advantage	Limitation
Chain-of-Thought	Simple; improves logic	Brittle; error accumulation
Tree-of-Thought	Multiple hypotheses	High compute overhead
Meta-CoT	Models reasoning process	Specialized tuning
RLVR-CoT	Emergent self-correction	Language mixing
ArgLLM-CoT	Formal contestability	Requires symbolic grounding

pipeline. At step t, the model generates candidate s_t with justifications. A symbolic parser extracts claims and relations to construct the local QBAF. Dialectical strength is computed via gradual semantics (e.g., DF-QuAD (Freedman et al., 2025)). If $\sigma(s_t) < \tau$, the step is rejected and the model regenerates with attacking arguments as context. **Epistemic status:** This pipeline is demonstrated for factual claim verification (Freedman et al., 2025). Whether it scales to multi-step mathematical or legal reasoning remains conjectural (gap G5).

4.2 VeriCoT: Neuro-Symbolic Chain-of-Thought Validation

VeriCoT (Feng et al., 2025) translates reasoning steps into First-Order Logic (FOL) for symbolic verification. Three scaling challenges arise: **(i) Chain length**: VeriCoT operates on conjunctive fragments keeping per-step verification in polynomial time, scaling linearly. **(ii) Uncertainty**: when premises are probabilistic, standard FOL is insufficient; probabilistic extensions (e.g., Markov Logic Networks) introduce approximate inference. **(iii) Partially structured data**: verification against free-text requires neural extraction that may itself introduce errors. **Epistemic status:** (i) is established for chains up to 15 steps (Feng et al., 2025); (ii)–(iii) lack satisfactory solutions. Integration with neuro-symbolic temporal reasoning (Liang et al., 2025) enables handling temporal constraints in legal and medical domains.

Verification scope. We envision *risk-stratified* verification: (i) **high-risk decisions** undergo full QBAF + VeriCoT verification of every step; (ii) **medium-risk reasoning** verifies only steps introducing new factual claims; (iii) **low-risk generation** uses sampled verification. This bounds the overhead of G5 to a level proportional to the decision risk.

Faithfulness vs. post-hoc rationalization. Externalizing reasoning into arguments improves *contestability* but does not guarantee *faithfulness*. The QBAF represents the model's *stated* reasoning, not its causal internal process. Three partial mitigations exist: *consistency checks* (detecting when stated arguments contradict outputs), *mechanistic probes* (verifying neural activations correlate with argumentative structure (Adadi and Berrada, 2018)), and *adversarial testing* (modifying premises to check output sensitivity). **Epistemic status:** Only consistency checks are currently implementable at scale.

The result is *contestable reasoning chains* that can be audited and challenged, even if full faithfulness guarantees remain open.

Gap Summary (Verification Dimension): ArgLLMs and VeriCoT address G3–G4 by externalizing reasoning into verifiable structures. Remaining: (G5) computational overhead of symbolic verification, and (G6) grounding arguments in distributed knowledge sources.

5 Federated Learning: The Engine of Sovereignty

FL addresses G1 (distributed routing), G2 (non-IID training), and G6 (distributed knowledge grounding).

5.1 The Non-IID Challenge and Coupled Behavior

Real-world federated data is **Non-IID** (de Ocáriz Borde, 2025). Standard FedAvg (McMahan et al., 2017) fails when $P_i(x, y) \neq P_j(x, y)$. The global FL objective is:

$$\min_{\theta} \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta) \tag{4}$$

where n_k is the number of samples at client k and $\mathcal{L}_k(\theta) = \mathbb{E}_{(x,y)\sim P_k}[\ell(\theta; x, y)]$. **Coupled Behavior Analysis** (de Ocáriz Borde, 2025) models client relationships through a coupling matrix $\mathbf{C}$:

$$C_{ij} = \text{sim}(P_i, P_j) = 1 - D_{\text{JS}}(P_i \| P_j) \tag{5}$$

with D_{JS} denoting the Jensen-Shannon divergence. This coupling enables **Personalized FL**:

$$\min_{\{\theta_k\}} \sum_{k=1}^{K} \left[\mathcal{L}_k(\theta_k) + \lambda \sum_{j \neq k} C_{kj} \|\theta_k - \theta_j\|^2 \right] \quad (6)$$

encouraging similar clients to learn similar models while dissimilar ones specialize. This aligns with the Sovereign MoE (Section 2): the router learns global patterns while sovereign experts capture local distributions. In our architecture, each expert e_k corresponds to a cluster of coupled clients, with the router activating experts based on input characteristics correlating with coupling structure:

$$G(x) = \text{TopK}\left(\text{softmax}\left(\mathbf{W}_g \cdot \phi(x)\right), k\right) \quad (7)$$

Epistemic status: The coupling matrix and personalized FL objective are established (de Ocáriz Borde, 2025)(McMahan et al., 2017). The MoE routing integration (Equation 4) is our extrapolation—mathematically consistent but empirically unvalidated. Whether the router can learn coupling-aligned representations and whether cross-institution expert activation is practically feasible remain open (gap G1).

5.2 Autonomous FL Synthesis and Privacy

The **Helmsman** framework (Li et al., 2025) uses multi-agent LLM collaboration to automatically synthesize FL architectures. For privacy, FL combines with **Differential Privacy** or **Generative FL** (Zhang et al., 2021), where clients share synthetic data rather than raw gradients, preserving privacy while learning causal structures. This enables training on sensitive data (medical imaging (Hartmann et al., 2024)) without centralization while ensuring GDPR compliance.

6 Knowledge Grounding: Virtual Knowledge Graphs

Gap G6 requires verifying claims against authoritative data without centralizing it. Virtual Knowledge Graphs (VKGs) (Xiao et al., 2019) address this: data remains in its original source while a conceptual ontology layer maps it to a graph structure *without moving it*. The reasoning agent generates SPARQL queries; the Ontop system (Xiao et al., 2019) translates these into SQL executed *locally* at each source; results are aggregated and returned—the model grounds reasoning *without ingesting raw private data*.

Centrality of unstructured knowledge (G8). Much evidence in law, medicine, and administration is unstructured text. G8 is a *critical path* issue. We envision a hybrid architecture: VKGs handle structured data with formal guarantees, while federated RAG (Lewis et al., 2020) handles unstructured data. The verification layer (Section 4) applies uniformly, with provenance metadata indicating evidence quality. **Epistemic status:** This hybrid is conjectural; no current system integrates federated VKG and federated RAG with unified verification.

When an ArgLLM generates a claim, it queries the VKG for supporting evidence, grounding the QBAF in authoritative data that is traceable, auditable, and accessed without leaving its source jurisdiction.

Gap Summary (Knowledge): VKGs address G6 for structured data but not: (G7) ontology mappings at scale—cross-jurisdiction terminological heterogeneity ("physician" vs. "Arzt" vs. "médecin"), evolving regulations, and absent governance for mapping conflicts; (G8) unstructured knowledge, where RAG (Lewis et al., 2020) addresses retrieval but lacks VKG's formal guarantees.

7 Embodied, Cultural, and Green AI

Beyond the core challenge of verifiable reasoning over distributed text, three additional dimensions shape the practical viability of sovereign AI systems. **Cognitive Robotics:** Erdem et al. (Ugur et al., 2025) integrate Answer Set Programming with motion planning, enabling formally verified robot actions—relevant for industrial AI over sovereign manufacturing data. **Cultural Alignment:** Cross-linguistic research (Blasi et al., 2022) highlights that models trained on English-dominated corpora exhibit biases affecting European language performance; affective reasoning modules can evaluate cultural appropriateness. **Green AI:** Inference-time scaling is energy-intensive; AutoML techniques (Tornede et al., 2023), adaptive compute allocation (Snell et al., 2024)(Zhu et al., 2025), early exit mechanisms (Schuster et al., 2022), and learning when to engage deliberative planning (Paglieri et al., 2025) address sustainability.

8 The EuReason Framework

The preceding sections identified eight gaps (G1–G8). EuReason organizes these into three lay-

ers (Figure 1): **Layer 1 (Sovereign Data)**: VKGs (Xiao et al., 2019) map heterogeneous sources to a shared ontology with GDPR compliance. **Layer 2 (Federated Learning)**: Helmsman-style agents (Li et al., 2025) handle Non-IID distributions with coupled learning (de Ocáriz Borde, 2025) and generative privacy (Hartmann et al., 2024). **Layer 3 (Neuro-Symbolic Reasoner)**: sparse MoE generator, QBAF verifier (Freedman et al., 2025), spatial grounding (Ugur et al., 2025), and cultural alignment.

8.1 Process Flow: Medical Compliance Example

Consider: *"Analyze compliance of this medical device across EU member states."* The Reasoner decomposes via System 2 deliberation. VKG queries dispatch to sovereign legal/medical databases; results aggregate via FL fabric. The MoE generator drafts analysis; the QBAF checks logical consistency while VeriCoT (Feng et al., 2025) validates against EU law. If attacks succeed, backtracking is triggered. Output: a verified, contestable report traceable to authoritative sources.

The dependency structure: G6 requires G7; G3 depends on G4; G1–G2 are prerequisites for G6. Thus ontology construction (G7) and trust infrastructure (G1) should be prioritized. Figure 2 visualizes this directed dependency graph.

8.2 Inter-Layer Interfaces and Failure Modes

Layer 1 → 2 (Data → Learning): VKG SPARQL endpoints serve the FL fabric. *Failure*: ontology drift when sources update schemas without updating mappings. *Mitigation*: schema versioning with change-detection hooks.

Layer 2 → 3 (Learning → Reasoning): FL provides trained expert weights; MoE router selects experts. *Failure*: stale experts from distribution shift without re-training. *Mitigation*: confidence decay functions reducing selection probability over time.

Layer 3 internal (Generator → Verifier): MoE generates; QBAF accepts/rejects. *Failure*: verification deadlock from overly strict τ or insufficient evidence. *Mitigation*: adaptive threshold relaxation with human escalation after k rejections.

8.3 Evaluation Framework

We propose concrete metrics for assessing progress on each gap:

9 Related Work

Our work draws on neuro-symbolic AI (Freedman et al., 2025; Colelough and Regli, 2025; Feng et al., 2025; Liang et al., 2025), federated learning (McMahan et al., 2017; de Ocáriz Borde, 2025; Li et al., 2025; Hartmann et al., 2024), inference-time scaling (Snell et al., 2024; Wu et al., 2025; Zhu et al., 2025; Paglieri et al., 2025), and ontology-based data access (Xiao et al., 2019). These build on transformer foundations (Vaswani et al., 2017; Brown et al., 2020; Devlin et al., 2019; Shazeer et al., 2017) and reasoning advances (Wei et al., 2022; Xiang et al., 2025; DeepSeek-AI, 2025). Trust and governance (Adadi and Berrada, 2018; Crockett et al., 2023; Tidjon and Khomh, 2022; Russell, 2019; Huang et al., 2024; Tornede et al., 2023) address deployment challenges. We analyze how these streams interact and where integration gaps remain.

9.1 Neighboring Paradigms

RAG (Lewis et al., 2020) grounds outputs in retrieved documents but provides no formal reasoning guarantees and assumes centralized retrieval. Our VKG-based grounding preserves sovereignty; our verification provides contestability. **Tool-augmented LLMs** (Schick et al., 2024) address capability gaps via external APIs—complementary, as tool use extends what models *can do* while our framework addresses whether outputs *can be trusted*. **Modular pipelines** (DSPy, LangChain) share our decomposition emphasis but lack formal verification and assume centralized orchestration. **Conventional trustworthy AI** (Crockett et al., 2023)(Huang et al., 2024) focuses on individual model properties; our framework extends trust to *distributed multi-component systems* with architectural failure modes (stale experts, ontology drift, verification deadlock).

10 Conclusion: Research Gaps and Priorities

This paper asked a single question: *how can AI systems reason verifiably over knowledge that must remain distributed?* By analysing recent advances across sparse MoE architectures, inference-time reasoning, neuro-symbolic verification, and federated learning, we identified eight research gaps and showed that they form a directed dependency structure. The "Jagged Intelligence" of current models (Dell'Acqua et al., 2023)—excelling at

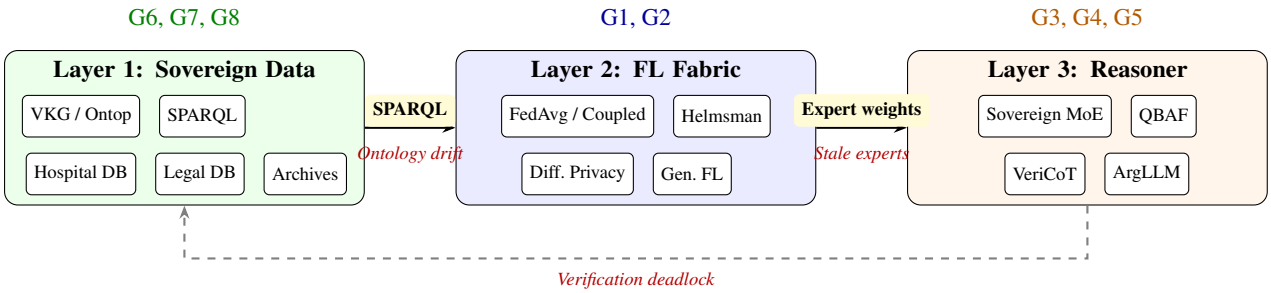

Figure 1: EuReason three-layer architecture. Solid arrows show forward data flow with inter-layer interfaces (yellow labels); the dashed arrow shows the verification feedback loop. Failure modes (red, italic) and associated research gaps are annotated per layer.

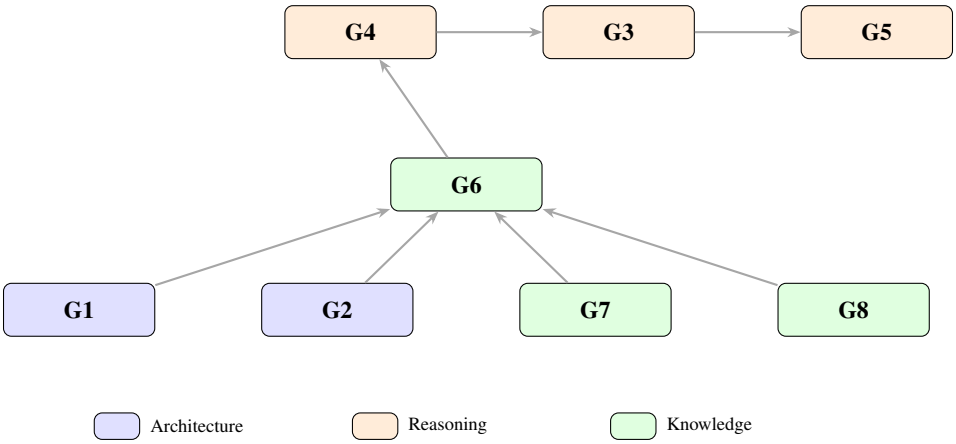

Figure 2: Directed dependency graph of research gaps G1–G8. Arrows indicate "is prerequisite for." Architecture (G1–G2) and knowledge foundations (G7–G8) converge on G6, which gates the reasoning chain G4→G3→G5.

benchmarks while failing simple reasoning (Chollet et al., 2024)—suggests the solution lies in verification, grounding, and distributed learning rather than larger models. The **EuReason** framework structures how architectural evolution (dense to sparse MoE), reasoning advances (System 1 to System 2), and governance models (centralised to federated) relate as a directed dependency graph. Five priorities emerge from the eight identified gaps: Federated Expert Routing (G1–G2), Knowledge Integration (G6–G8), Argumentative Verification (G3–G4), Efficient Neuro-Symbolic Methods (G5), and Green and Cultural Alignment—together defining a research agenda toward reasoning systems where verifiability, data sovereignty, and auditability are first-class architectural constraints.

Impact. If the gaps identified here are resolved, three practical consequences follow. First, organisations in regulated sectors—hospitals, courts, public agencies—could deploy reasoning systems that satisfy GDPR and EU AI Act requirements *by design* rather than through post-hoc compliance layers. Second, the dependency graph (Figure 2) offers the research community a concrete prioritisation: investments in trust infrastructure (G1) and ontology construction (G7) will unblock the largest number of downstream advances. Third, contestable, source-traceable reasoning chains would shift the trust model from "trust the model" to "audit the argument," enabling human oversight at the level of individual reasoning steps rather than aggregate outputs.

Limitations

This paper is a position and survey paper that synthesizes existing literature rather than presenting new empirical results. The proposed EuReason framework is conceptual and has not been implemented or evaluated end-to-end; while we specify inter-layer interfaces, failure modes, and evaluation metrics, these remain proposals rather than validated designs. The European sovereignty perspective, while motivated by regulatory and strategic considerations, may not generalize to other geopolitical contexts. The rapid pace of AI development means that some technological assessments may

Table 3: Proposed Evaluation Metrics for EuReason Gaps

Gap	Metric
G1	Query latency ratio: federated vs. centralized routing (target: <3x overhead)
G2	Accuracy gap between federated and centralized training on non-IID benchmarks
G3	Faithfulness score: correlation between stated reasons and causal attributions
G4	Grounding precision: fraction of verified claims traceable to authoritative sources
G5	Verification overhead: wall-clock time ratio of verified vs. unverified inference
G6	Cross-source recall: fraction of relevant distributed facts retrieved
G7	Ontology coverage: percentage of domain terms mapped across n jurisdictions
G8	Hybrid retrieval F1: combined structured (VKG) and unstructured (RAG) retrieval

become outdated quickly.

Scope of comparison. While we compare our approach with RAG, tool-augmented LLMs, modular pipelines, and conventional trustworthy AI (Section 9), other relevant paradigms—such as neurosymbolic program synthesis, world models, and constitutional AI—receive limited treatment. A comprehensive comparison across all neighboring paradigms exceeds the scope of a single position paper.

Reliance on recent preprints. A significant portion of our argument depends on preprints and rapidly evolving claims about frontier-model reasoning (2024–2026). While we prioritize work from peer-reviewed venues (AAAI, NeurIPS, ICLR, ACL) where available, some key results—particularly on inference-time scaling and federated agent synthesis—are available only as preprints. We have marked the epistemic status of claims throughout; conclusions depending on preprint results are correspondingly less stable.

Alternative paths to compliance. The European regulatory context (GDPR, EU AI Act) motivates our technical stack, but alternative paths to compliant and trustworthy AI exist. Centralized systems with strong access controls and audit logging can also satisfy regulatory requirements without federated architectures. Our argument is not that EuReason is the *only* path to compliance, but that it offers advantages—data minimization by design, distributed trust, and formal contestability—that centralized approaches achieve only through additional engineering layers. We acknowledge that for organizations with existing centralized infrastructure and established governance, conventional approaches may be more practical in the near term.

Maturity of Component Technologies. The EuReason framework integrates components at varying levels of maturity, and we assess the key limitations of each. Neuro-symbolic verification approaches like VeriCoT (Feng et al., 2025) can detect logically inconsistent reasoning chains by translating chain-of-thought steps into First-Order Logic, but gap G5 (verification overhead) conflicts with real-time inference requirements. Virtual Knowledge Graphs (Xiao et al., 2019) ground neural outputs in authoritative data through semantic mappings rather than document ingestion (Lewis et al., 2020), but gaps G7–G8 (ontology construction at European scale and unstructured knowledge handling) limit their applicability. Argumentative AI frameworks (Freedman et al., 2025) enable contestability by structuring claims with explicit evidence and counterarguments—addressing the EU AI Act (European Parliament and Council of the European Union, 2024)—but require symbolic grounding infrastructure that does not yet exist at scale. The Helmsman framework (Li et al., 2025) demonstrates that multi-agent systems can synthesize federated learning architectures, but gap G1 (distributed routing) requires trust infrastructure not yet available at European scale. These unresolved integration challenges define the research agenda that EuReason structures.

Ethical Considerations

This work discusses AI sovereignty and federated learning in the context of European values including privacy, transparency, and human oversight. We acknowledge that sovereignty-focused AI development could inadvertently create technological barriers. The medical compliance example used throughout is illustrative and does not constitute medical advice or validated clinical methodology.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.

Meta AI. 2025. Llama 4: Open multimodal foundation models. Technical report, Meta Platforms.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

François Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2024. ARC prize 2024: Technical report. *Preprint*, arXiv:2412.04604.

Brandon C. Colelough and William Regli. 2025. Neuro-symbolic ai in 2024: A systematic review. *Preprint*, arXiv:2501.05435.

Keeley Crockett, Edwin Colyer, Luciano Gerber, and Annabel Latham. 2023. Building trustworthy AI solutions: A case for practical solutions for small businesses. *IEEE Transactions on Artificial Intelligence*, 4(6):1456–1469.

Haitz Sáez de Ocáriz Borde. 2025. Sovereign ai vs ai polyglots. *SSRN Electronic Journal*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5639470.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI act). Official Journal of the European Union L 2024/1689. Full applicability: August 2026.

Yu Feng, Nathaniel Weir, Kaj Bostrom, Sam Bayless, Darion Cassel, Sapana Chaudhary, Benjamin Kiesl-Reiter, and Huzefa Rangwala. 2025. Vericot: Neuro-symbolic chain-of-thought validation via logical consistency checks. *Preprint*, arXiv:2511.04662.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative large language models for explainable and contestable claim verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14930–14939.

Yale Hartmann, Jonah Klöckner, Lucas Deichsel, Rinu Paul, and Tanja Schultz. 2024. Gait parameter estimation from a single privacy preserving depth sensor. In *Proceedings of BIOSIGNALS*, pages 637–645.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, and 1 others. 2024. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):176.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Haoyuan Li, Mathias Funk, and Aaqib Saeed. 2025. Helmsman: Autonomous synthesis of federated learning systems via collaborative llm agents. *Preprint*, arXiv:2510.14512.

Feng Liang, Weixin Zeng, Runhao Zhao, and Xiang Zhao. 2025. Nestr: A neuro-symbolic abductive framework for temporal reasoning in large language models. *Preprint*, arXiv:2512.07218. Accepted by AAAI 2026.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

OpenAI. 2024. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2026-01-02.

Davide Paglieri, Jelena Luketina, Ying Fan, and Karl Cobbe. 2025. Learning when to plan: Efficiently allocating test-time compute for LLM agents. *arXiv preprint arXiv:2509.03581*.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The curse of recursion: Training on generated data makes models forget. *Nature*, 631:755–760.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Lionel Nganyewou Tidjon and Foutse Khomh. 2022. The different faces of ai ethics across the world: A principle-implementation gap analysis. *Preprint*, arXiv:2206.03225.

Tanja Tornede, Alexander Tornede, Jonas Hanselle, Felix Mohr, Marcel Wever, and Eyke Hüllermeier. 2023. Towards green automated machine learning: Status quo and future directions. *Journal of Artificial Intelligence Research*, 77:427–457.

Emre Ugur, Alper Ahmetoglu, Yukie Nagai, Tadahiro Taniguchi, Matteo Saveriano, and Erhan Oztop. 2025. Neuro-symbolic robotics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *Preprint*, arXiv:2408.00724.

Violet Xiang and 1 others. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought. *arXiv preprint arXiv:2501.04682*.

Guohui Xiao, Linfang Ding, Benjamin Cogrel, and Diego Calvanese. 2019. Virtual knowledge graphs: An overview of systems and use cases. *Data Intelligence*, 1(3):201–223.

Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems*, 216:106775.

Yinglun Zhu and 1 others. 2025. Strategic scaling of test-time compute: A bandit learning approach. *arXiv preprint arXiv:2506.12721*.