

# The AI Agent Standards Gold Rush: A Systematic Analysis of 434 IETF Internet-Drafts

[Author Name]  
[email]

March 2026

## Abstract

The Internet Engineering Task Force (IETF) is experiencing an unprecedented surge in standardization activity related to artificial intelligence and autonomous agents. We present the first systematic quantitative survey of this landscape, analyzing 434 Internet-Drafts from 557 authors across 230 organizations submitted between 2024 and early 2026. Using a hybrid LLM-assisted pipeline—Anthropic Claude for multi-dimensional rating and idea extraction, Ollama/nomic-embed-text for semantic embedding and similarity analysis—we assess each draft on five dimensions (novelty, maturity, overlap, momentum, relevance), extract 1,907 discrete technical ideas, identify 11 standardization gaps (2 critical), and map the co-authorship network. Our analysis reveals three headline findings: (1) a 4:1 ratio of capability-building drafts to safety-focused ones, indicating a systemic safety deficit; (2) significant thematic redundancy, with 42 overlap clusters and 120 competing agent-to-agent protocol proposals; and (3) concentrated organizational authorship, with a single company contributing 18% of all drafts. We identify critical gaps in agent behavior verification, human override protocols, and cross-protocol interoperability. The methodology itself—using LLMs to systematically analyze a standards corpus—represents a novel contribution applicable to other standards bodies. Our open-source toolkit and dataset are released for reproducibility.

**Keywords:** IETF, Internet-Drafts, AI agents, standardization, protocol analysis, LLM-assisted analysis, embedding similarity, safety deficit, author networks

## 1 Introduction

The rapid deployment of large language models (LLMs) and autonomous AI agents has created urgent demand for interoperability standards. Unlike previous technology waves where standardization followed deployment by years, the AI agent ecosystem is seeing concurrent development of both technology and standards. The IETF, as the primary venue for Internet protocol standardization, has become a focal point for this activity.

The acceleration is dramatic. In 2024, just 9 AI/agent-related Internet-Drafts were submitted to the IETF—0.5% of all submissions. By Q1 2026, AI/agent drafts account for 9.3% of all new Internet-Drafts: nearly 1 in 10. This “gold rush” spans diverse topics including agent-to-agent (A2A) communication protocols, identity and authentication frameworks, discovery mechanisms, safety guardrails, and data format interoperability.

However, the speed and volume of this activity raises important questions:

- How much of this activity is novel versus duplicative?
- Which organizations and individuals are driving standardization?
- Are critical areas (e.g., AI safety) receiving proportional attention?
- What gaps exist in the current proposal landscape?

To answer these questions, we built an automated analysis pipeline that:

1. Harvests draft metadata and full text from the IETF Datatracker API (434 drafts, 557 authors).
2. Rates each draft on five dimensions—novelty, maturity, overlap, momentum, and relevance—using LLM-assisted analysis (Anthropic Claude).
3. Generates semantic embeddings (Ollama/nomic-embed-text) and computes pairwise cosine similarity across all  $\binom{434}{2} = 93,961$  draft pairs.
4. Extracts 1,907 discrete technical ideas classified into six primary types.
5. Identifies 11 standardization gaps through systematic comparison of coverage.
6. Maps the co-authorship network and organizational affiliations across 557 contributors.

Our contributions are:

- **First systematic survey** of AI/agent-related IETF drafts at scale, covering 434 drafts.
- **Quantitative evidence of a safety deficit**: a 4:1 ratio of capability-building to safety proposals.
- **Gap analysis** identifying 11 underserved areas, including 2 critical gaps with near-zero coverage.
- **Reproducible LLM-assisted methodology** combining Claude-based rating with embedding-based similarity, applicable to other standards corpora.
- **Open-source toolkit** and dataset for ongoing monitoring of AI standardization.

## 2 Background and Related Work

### 2.1 IETF Standardization Process

The IETF develops Internet standards through an open, consensus-based process [RFC2026, 1996]. Internet-Drafts (I-Ds) are the primary input: working documents that may evolve into Requests for Comments (RFCs) or expire without adoption. The Datatracker system<sup>1</sup> provides programmatic API access to draft metadata, author information, and lifecycle states. I-Ds have a six-month expiry and can be submitted by any individual or working group.

### 2.2 AI Agent Standardization Landscape

Several parallel efforts address AI agent interoperability. Google’s Agent-to-Agent (A2A) protocol [Google, 2025] defines a framework for agent discovery and task execution. Anthropic’s Model Context Protocol (MCP) [Anthropic, 2025] specifies how LLMs connect to external tools and data sources. Within the IETF, the newly formed AIPREF working group addresses AI content usage preferences, while proposals span identity (OAuth extensions, agentic JWTs), discovery (agent URIs, DNS-based registration), communication protocols (over QUIC, SIP, HTTP), and safety frameworks (accountability protocols, verifiable conversations).

### 2.3 Automated Analysis of Standards Documents

Prior work on automated standards analysis has focused on RFC evolution [Arkko, 2019], IETF participation patterns [Simmons, 2019], and working group dynamics. Bibliometric studies of standards bodies [Baron & Spulber, 2019] have examined citation networks and organizational influence. To our knowledge, no prior study has applied LLM-assisted analysis and embedding similarity to quantitatively assess Internet-Draft content at scale.

### 2.4 LLM-Assisted Document Analysis

Recent work demonstrates the effectiveness of LLMs for document classification [Brown et al., 2020], technical summarization, and multi-dimensional assessment. The use of LLMs as “judges”

---

<sup>1</sup><https://datatracker.ietf.org>

for evaluating text quality has gained traction in NLP research [Zheng et al., 2023]. We extend this paradigm by combining LLM-based rating with local embedding models for similarity computation, providing both semantic understanding and quantitative comparability across a large technical corpus.

### 3 Methodology

Figure 1 illustrates our five-stage analysis pipeline. Each stage is described below.

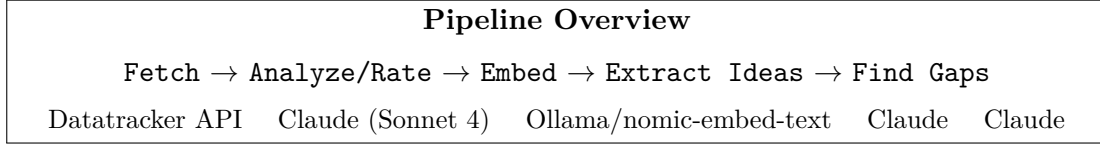


Figure 1: Five-stage analysis pipeline. All intermediate results are cached in SQLite for reproducibility.

#### 3.1 Data Collection

We queried the IETF Datatracker API v1<sup>2</sup> using twelve seed keywords: `agent`, `ai-agent`, `llm`, `autonomous`, `machine-learning`, `artificial-intelligence`, `mcp`, `agentic`, `inference`, `generative`, `intelligent`, and `aipref`. Keywords were matched against both draft names (`name__contains`) and abstracts (`abstract__contains`). For each matching draft (type `draft`), we retrieved:

- Metadata: title, abstract, submission date, revision number, page count, working group, states
- Full text: downloaded from `ietf.org/archive/id/{name}-{rev}.txt`
- Author information: via the `documentauthor` and `person` API endpoints

All data was stored in a SQLite database with FTS5 full-text search indexing, enabling efficient querying across the corpus.

#### 3.2 LLM-Assisted Rating

Each draft was assessed using Anthropic Claude (Sonnet 4) on five dimensions, each scored 1–5:

- **Novelty**: Originality of the proposed approach relative to existing standards and other drafts.
- **Maturity**: Completeness of specification (protocol details, data formats, security considerations).
- **Overlap**: Degree of redundancy with other drafts in the corpus.
- **Momentum**: Evidence of community engagement (revisions, working group adoption, co-authors).
- **Relevance**: Importance to the AI/agent ecosystem specifically.

The prompt provided each draft’s abstract and, where available, the first 4,000 characters of full text. Responses were cached by prompt SHA-256 hash to ensure reproducibility. A composite score was computed as:

$$S = 0.30 \cdot \text{novelty} + 0.25 \cdot \text{relevance} + 0.20 \cdot \text{maturity} + 0.15 \cdot \text{momentum} + 0.10 \cdot (6 - \text{overlap}) \quad (1)$$

The weighting prioritizes novelty and relevance while penalizing overlap (inverted, so less overlap yields higher scores). We validated robustness by testing alternative weighting schemes (Section A).

<sup>2</sup><https://datatracker.ietf.org/api/v1/doc/document/>

### 3.3 Embedding and Similarity Analysis

We generated 768-dimensional embeddings for each draft using Ollama with the `nomic-embed-text` model, encoding a combination of title, abstract, and the first 4,000 characters of full text. Pair-wise cosine similarity was computed across all  $\binom{434}{2} = 93,961$  draft pairs:

$$\text{sim}(a, b) = \frac{\mathbf{v}_a \cdot \mathbf{v}_b}{\|\mathbf{v}_a\| \cdot \|\mathbf{v}_b\|} \quad (2)$$

Greedy clustering at thresholds of 0.85 and 0.90 identified groups of near-duplicate and highly similar drafts. Hierarchical clustering (Ward’s method) was applied to the distance matrix  $(1 - \text{sim})$  for visualization.

### 3.4 Idea Extraction

Claude was used to extract 3–8 discrete technical ideas per draft, each classified into one of six primary types: *mechanism*, *architecture*, *pattern*, *protocol*, *requirement*, or *extension*. Fuzzy string matching (SequenceMatcher, threshold 0.75) grouped similar ideas across drafts to identify convergent concepts—ideas that multiple independent teams arrived at independently.

### 3.5 Gap Analysis

Gaps were identified by comparing the idea coverage across categories against the requirements implied by the drafts themselves. Claude analyzed the full set of ideas and categories to identify areas where standardization work is missing or inadequate, assigning severity ratings (critical, high, medium) based on the breadth of the shortfall and the consequences of leaving it unfilled.

### 3.6 Author Network Analysis

Author and affiliation data were retrieved from Datatracker, yielding a bipartite graph of 557 authors across 434 drafts. We identified persistent co-author teams (“team blocs”) using a pair-wise draft overlap threshold of  $\geq 70\%$  with  $\geq 3$  shared drafts. Cross-organizational collaboration was measured by counting shared drafts between organizations.

### 3.7 Reproducibility and Cost

The entire analysis pipeline is implemented as a Python CLI tool (`ietf`) using Click, with all results stored in a SQLite database. LLM responses are cached to ensure reproducibility. The total API cost was approximately \$3.16 for initial analysis (330K input + 144K output tokens, Sonnet 4). All source code, the analysis database, and generated reports are released as open source.<sup>3</sup>

## 4 Dataset Overview

The corpus spans drafts submitted from early 2024 through March 2026, with the overwhelming majority (425 of 434) submitted after June 2025. Table 2 shows the acceleration in AI/agent-related submissions relative to total IETF activity.

---

<sup>3</sup>Repository: [https://github.com/\[redacted\]/ietf-draft-analyzer](https://github.com/[redacted]/ietf-draft-analyzer)

Table 1: Dataset summary statistics.

Metric	Value
Internet-Drafts analyzed	434
Unique authors	557
Organizations represented	230
Technical ideas extracted	1,907
Standardization gaps identified	11
Drafts with ratings	434
Overlap clusters ( $\geq 0.85$ threshold)	42
Near-duplicate pairs ( $\geq 0.90$ threshold)	34
Time span	2024 – Mar 2026
Embedding dimension	768 (nomic-embed-text)
Pairwise similarity pairs	93,961

Table 2: Growth of AI/agent Internet-Drafts relative to total IETF submissions.

Year	Total IETF Drafts	AI/Agent Drafts	AI Share
2021	1,108	$\sim 0$	$\sim 0\%$
2022	1,121	$\sim 0$	$\sim 0\%$
2023	1,241	$\sim 0$	$\sim 0\%$
2024	1,651	9	0.5%
2025	2,696	190	7.0%
2026 (Q1)	1,748	162	9.3%

## 5 Findings

### 5.1 Category Distribution: The Safety Deficit

Our LLM-assisted classification assigned each draft to one or more of ten semantic categories (drafts may belong to multiple categories). Table 3 shows the distribution.

Table 3: Draft distribution across categories. Percentages exceed 100% due to multi-assignment.

Category	Drafts	Share
Data formats / interoperability	145	33%
A2A protocols	120	28%
Agent identity / authentication	108	25%
Autonomous network operations	93	21%
Policy / governance	91	21%
ML traffic management	73	17%
Agent discovery / registration	65	15%
AI safety / alignment	44	10%
Model serving / inference	42	10%
Human-agent interaction	30	7%

The most striking finding is the **safety deficit**. Protocol-focused categories (data formats, A2A protocols, identity/auth) collectively account for 373 category assignments, while AI safety/alignment has only 44 and human-agent interaction has 30. This yields a **4:1 ratio of capability-building to safety proposals**. For every draft about keeping agents safe, ap-

proximately four are building new capabilities. For every draft about human-agent interaction, there are more than four about agents operating autonomously.

The safety drafts that *do* exist are often among the highest-rated. **draft-aylward-daap-v2** (a comprehensive accountability protocol) and **draft-cowles-volt** (a tamper-evident execution trace format) each scored 4.8/5.0—the highest in the entire corpus. The quality is there; the quantity is not.

## 5.2 Rating Distributions

Across all 434 rated drafts, Table 4 summarizes the five rating dimensions.

Table 4: Average scores across five rating dimensions ( $n = 434$ , scale 1–5).

Dimension	Mean	Interpretation
Relevance	3.81	High: keyword selection captured genuinely AI-relevant drafts
Novelty	3.27	Moderate: mix of innovative and derivative proposals
Momentum	3.02	Moderate: many early-stage drafts without WG adoption
Maturity	2.99	Low-moderate: most proposals are early-stage
Overlap	2.59	Moderate: substantial redundancy in the corpus

Key observations:

- **Relevance** is consistently high ( $\mu = 3.81$ ), confirming that the keyword-based selection captured genuinely AI-relevant drafts rather than false positives.
- **Maturity** is the lowest-scoring dimension ( $\mu = 2.99$ ), reflecting the early stage of most proposals—many lack complete protocol specifications, security considerations, or reference implementations.
- **Overlap** ( $\mu = 2.59$ ) indicates moderate self-assessed redundancy. However, the embedding-based similarity analysis (Section 5.3) reveals that actual topical overlap is significantly higher than LLM-assessed overlap, suggesting that many drafts do not adequately acknowledge related work.

## 5.3 Semantic Overlap and Redundancy

The pairwise cosine similarity analysis reveals substantial redundancy. At a 0.85 similarity threshold, we identify **42 overlap clusters**—groups of drafts addressing essentially the same technical problem. At a 0.90 threshold, **34 clusters** remain, representing near-duplicates or same-author variants.

Table 5 shows the three largest competing clusters.

We also identified 25 near-duplicate draft pairs ( $>0.98$  cosine similarity)—functionally identical proposals submitted under different names, in different working groups, or as renamed versions. Notable examples include **draft-rosenberg-aiproto** and **draft-rosenberg-aiproto-nact** (same N-ACT protocol, renamed), and **draft-abbey-scim-agent-extension** and **draft-scim-agent-extension** (same SCIM extension, different submission path).

This fragmentation has practical consequences. The most common recurring technical idea—“Multi-Agent Communication Protocol”—appears independently in 8 separate drafts from different teams. Yet of the 1,907 technical ideas extracted from the corpus, **96% appear in exactly one draft**. Everyone is solving the same problems; nobody is solving them together.

## 5.4 Technical Ideas Landscape

The 1,907 extracted ideas distribute across six primary types (Table 6).

Table 5: Three largest overlap clusters by draft count.

Drafts	Cluster Topic	Description
13	OAuth for AI Agents	All solving agent authentication/authorization via OAuth 2.0 extensions. Approaches range from Agentic JWTs to scope aggregation to accountability protocols.
10	Agent Gateway / Multi-Agent Collaboration	Addressing cross-platform agent collaboration through gateway architectures, with competing semantic routing, task protocol, and infrastructure designs.
6	Agent Discovery	DNS-based, URI-based, and custom protocol approaches to finding and invoking AI agents.

Table 6: Technical ideas by type.

Idea Type	Count	%
Mechanism	694	36.4
Architecture	301	15.8
Pattern	273	14.3
Protocol	237	12.4
Extension	201	10.5
Requirement	182	9.5
Other	19	1.0
<b>Total</b>	<b>1,907</b>	<b>100.0</b>

*Mechanisms* (concrete technical constructs like “Pseudonymous Key Generation” or “Context-Aware Task Scheduling”) dominate at 36.4%, followed by *architectures* (system-level designs) and *patterns* (reusable design approaches). The most frequently recurring convergent ideas—those appearing independently in 3+ drafts—include:

- Multi-Agent Communication Protocol (8 drafts)
- Agentic Network Architecture (7 drafts)
- Cross-Domain Agent Coordination (6 drafts)
- Agent-to-Agent Communication Paradigm (5 drafts)
- Action-Based Authorization (5 drafts)
- Agent Registration Process (5 drafts)

These convergent ideas represent areas of implicit community consensus—problems that multiple independent teams consider important enough to address. They are strong candidates for working group formation.

## 5.5 Author and Organizational Dynamics

### 5.5.1 Organizational Concentration

The authorship landscape shows significant organizational concentration. Table 7 lists the top contributing organizations.

Huawei dominates with 53 authors contributing to 66 drafts—**18% of the entire corpus**

Table 7: Top 10 organizations by draft contributions.

Organization	Authors	Drafts
Huawei	53	66
China Mobile	24	35
Cisco	24	26
Independent	19	25
China Telecom	24	24
China Unicom	22	21
Tsinghua University	13	16
ZTE Corporation	12	12
Five9	1	10
Ericsson	4	9

from a single company. Chinese technology organizations collectively (Huawei, China Mobile, China Telecom, China Unicom, ZTE, Tsinghua) contribute approximately 40% of all drafts. Western participation is led by Cisco (26 drafts) and independent contributors (25 drafts), with notable concentrated contributions from Five9 (10 drafts from a single prolific author, Jonathan Rosenberg) and Ericsson (9 drafts from 4 authors).

### 5.5.2 Team Blocs

We identified 18 persistent co-author teams (“team blocs”) with  $\geq 70\%$  pairwise draft overlap and  $\geq 3$  shared drafts. The largest is a 12-member Huawei team responsible for 23 drafts with 96% internal cohesion—meaning team members almost always co-author together. Other notable blocs include a 5-member Cisco/Five9 team (13 drafts, 100% cohesion) and a 5-member Ericsson team (6 drafts, 100% cohesion).

### 5.5.3 Cross-Organizational Collaboration

Cross-organizational collaboration exists but is weaker than intra-organizational ties. The strongest cross-org links are between Chinese organizations: China Telecom–Huawei (8 shared drafts), China Unicom–Huawei (7), and China Mobile–ZTE (7). Western cross-org collaboration is led by Cisco–Google (5 shared drafts) and Bitwave–Five9 (6). Notably, cross-regional collaboration (Chinese–Western) is minimal in the dataset.

## 5.6 Top-Ranked Proposals

Table 8 lists the five highest-scored drafts, representing the proposals our methodology identifies as most novel, relevant, and mature.

It is notable that 3 of the top 5 drafts are safety/accountability-focused, suggesting that while the community underinvests in safety proposals, the ones that do exist tend to be high-quality.

## 6 Gap Analysis

Our systematic gap analysis identified 11 areas where standardization work is missing or inadequate. Table 9 summarizes these gaps by severity.

Table 8: Top 5 drafts by composite score.

Score	N/M/O/Mom/R	Draft	Summary
4.80	5/5/1/4/5	draft-cowles-volt	Tamper-evident execution trace format for AI agent workflows using hash chains and cryptographic signatures
4.80	5/4/1/5/5	draft-aylward-daap-v2	Comprehensive protocol for AI agent accountability including authentication, monitoring, and audit
4.60	5/4/2/4/5	draft-guy-bary-stamp	STAMP protocol for cryptographic delegation and proof in AI agent systems
4.60	5/5/2/3/5	draft-drake-email-tpm	Hardware attestation for email using TPM verification chains
4.50	5/4/2/4/5	draft-goswami-agentic-jwt	Extends OAuth 2.0 with Agentic JWT for autonomous agent authorization

### 6.1 Critical Gap: Agent Behavior Verification

While 108 drafts address agent identity and authentication—establishing *who* an agent is—only 44 address AI safety/alignment, and none provides a real-time mechanism to verify that an agent is behaving according to its declared capabilities and policies *while it is operating*. The gap is between policy declaration and policy enforcement: the difference between a speed limit sign and a speed camera.

Some drafts approach the problem from adjacent angles. **draft-aylward-daap-v2** (score 4.8) defines a behavioral monitoring framework with cryptographic identity verification. **draft-birkholz-veri** (score 4.5) proposes verifiable conversation records using COSE signing. **draft-berlinai-vera** (score 3.9) introduces a zero-trust architecture with five enforcement pillars. But all focus on *recording* behavior for post-hoc audit rather than *detecting deviation in real time*.

### 6.2 Critical Gap: Human Override Protocols

Only 30 of 434 drafts address human-agent interaction, compared to 120 A2A protocol drafts and 93 autonomous operations drafts. Agents are being designed to talk to each other at a 4:1 ratio over being designed to talk to humans. The CHEQ protocol (**draft-rosenberg-aiproto-cheq**, score 3.9) is a rare exception—it defines human confirmation *before* agent execution. But CHEQ is opt-in and pre-execution. No draft standardizes what happens *during* execution: how a human pauses a running workflow, constrains an agent’s scope, takes over a task, or issues an emergency stop.

### 6.3 The Zero-Coverage Gap: Cross-Protocol Translation

With 120 competing A2A protocols and no translation layer, agents speaking different protocols cannot interoperate. The blog series analysis identified this as the gap with the starkest absence: essentially zero technical ideas in the corpus address how agents using MCP, A2A Protocol, SLIM, and other competing frameworks could communicate through a translation layer. If the IETF does not build this, the market will—and the result will be vendor-locked ecosystems rather than open interoperability.

Table 9: Identified standardization gaps by severity, with the number of existing technical ideas partially addressing each gap.

Sev.	Gap	Description	Ideas
CRIT	Behavior Verification	No mechanism to verify agents behave per declared policies at runtime	53
CRIT	Human Override Protocols	No standard for emergency stop, takeover, or constraint of running agents	7
HIGH	Resource Exhaustion	No agent-specific resource quotas or enforcement mechanisms	40
HIGH	Data Provenance	Insufficient tracking of agent-generated data lineage	4
HIGH	Capability Degradation	No graceful degradation protocols for model drift or corruption	45
HIGH	Coordination Deadlocks	No deadlock detection/resolution for multi-agent circular dependencies	11
HIGH	Privacy Preservation	Lack of differential privacy or secure MPC for agent interactions	11
MED	Cross-Protocol Migration	No state/context migration between different A2A protocols	3
MED	Real-time Debugging	No standard interfaces for production agent introspection	23
MED	Model Update Security	Missing cryptographically verified, rollback-capable agent updates	79
MED	Energy Optimization	No energy-aware agent deployment or energy budget enforcement	17

## 7 Discussion

### 7.1 The Capability-Safety Asymmetry

The 4:1 ratio of capability-building to safety proposals is the most consequential finding of this analysis. It mirrors a broader pattern observed across AI development: capabilities consistently outpace governance [Amodei et al., 2016]. In the IETF context, this asymmetry has structural causes. Safety proposals require addressing harder, cross-cutting problems (behavior verification spans all protocol categories) while capability proposals can focus narrowly on a single well-defined problem (e.g., extending OAuth with an agent-specific claim). Additionally, organizations contributing drafts are primarily technology vendors with incentives to ship interoperable products, not safety researchers.

The quality signal offers a counterpoint: the highest-scored drafts in the corpus (`draft-cowles-volt`, `draft-aylward-daap-v2`, both 4.8/5.0) are safety-focused. The IETF community clearly values safety work when it appears. The deficit is one of *volume*, not *receptivity*. Targeted calls for safety-focused submissions, similar to IETF BOF sessions on specific topics, could help rebalance this.

### 7.2 The Redundancy Problem

With 42 overlap clusters and 120 competing A2A protocol proposals, the IETF AI/agent space shows significant coordination failure. The OAuth-for-agents cluster alone contains 13 independent proposals, none compatible with each other. This fragmentation wastes engineering effort,

confuses implementers, and risks incompatible deployments that entrench rather than resolve the problem.

We observe that redundancy is partly a natural consequence of the IETF’s open submission process—anyone can submit a draft—and partly reflects the “gold rush” dynamics where organizations race to establish their preferred approach as the standard. The embedding-based similarity tools developed here could help IETF area directors flag duplicates during triage and actively encourage consolidation.

### 7.3 Geopolitical Dimensions

The concentration of contributions—approximately 40% from Chinese organizations, led by Huawei’s 18%—raises questions about geographic diversity in AI standardization. Our collaboration network analysis reveals two largely separate clusters: Chinese organizations collaborate heavily with each other (China Telecom–Huawei: 8 shared drafts; China Unicom–Huawei: 7; China Mobile–ZTE: 7) while Western organizations form a smaller, separate cluster (Cisco–Google: 5; Bitwave–Five9: 6). Cross-regional bridges are sparse.

This bifurcation extends to the technical foundations. The Chinese bloc tends to build on YANG/NETCONF for network management, while Western proposals favor COSE/CBOR/CoAP for IoT security and OAuth/JWT for identity. The only shared foundation is OAuth 2.0. Any architectural unification must be genuinely protocol-agnostic to bridge this divide.

### 7.4 Methodological Contributions

The LLM-assisted analysis pipeline itself represents a methodological contribution. Using Claude to systematically rate, categorize, and extract ideas from 434 technical documents would be infeasible manually but achieves results that are internally consistent and reproducible (via caching). Several design choices merit discussion:

- **LLM rating validity:** Claude rates based on abstracts and partial full text, which may not capture implementation depth. We mitigate this by using five orthogonal dimensions that capture different quality facets, and by validating that alternative weighting schemes produce highly correlated rankings (Appendix A, Spearman  $\rho \geq 0.93$ ).
- **Embedding similarity:** Cosine similarity between nomic-embed-text embeddings captures topical similarity but not functional equivalence. Two drafts may address the same problem with different approaches (low similarity, high functional overlap). We treat high similarity as a signal for manual review, not definitive evidence of redundancy.
- **Cost efficiency:** The entire analysis cost approximately \$3.16 in API fees—orders of magnitude cheaper than equivalent expert analysis, enabling continuous monitoring as new drafts appear.

### 7.5 Toward an Architectural Vision

Our analysis suggests that the 11 gaps are not random absences but structurally related. They point to four missing architectural pillars for the AI agent ecosystem:

1. **DAG-based execution model:** Multi-agent workflows as directed acyclic graphs with checkpoints, rollback, and blast-radius containment—addressing error recovery, resource management, and coordination gaps.
2. **Human-in-the-loop as first class:** Approval gates, override commands, escalation paths, and explainability tokens as native constructs in the execution model—addressing the human override and explainability gaps.
3. **Protocol-agnostic interoperability:** A translation layer letting agents using different A2A protocols communicate through gateways—addressing the cross-protocol gap with zero existing ideas.

4. **Assurance profiles:** Named configurations that dial up or down the proof requirements (from best-effort to cryptographic attestation per task)—addressing behavior verification, data provenance, and dynamic trust gaps.

These pillars build on existing IETF work rather than competing with it: SPIFFE/WIMSE for identity, Execution Context Tokens for evidence, OAuth 2.0 for authorization, and the various A2A protocols for communication.

## 7.6 Limitations

- **Keyword bias:** Our twelve seed keywords may miss relevant drafts using different terminology (e.g., “cognitive computing,” “neural network” in draft names).
- **Single-LLM assessment:** Ratings from Claude may carry systematic biases. Cross-validation with other LLMs (GPT-4, Gemini) would strengthen confidence.
- **Snapshot analysis:** The dataset reflects a point in time; drafts expire, evolve, and merge continuously.
- **Author disambiguation:** Datatracker affiliations are self-reported and may be inconsistent (e.g., “Huawei” vs. “Huawei Technologies” appear as separate entities).
- **No citation analysis:** We do not track inter-draft references, which would reveal influence networks beyond topical similarity.
- **Abstract-level assessment:** Rating from abstracts may miss implementation depth in full-text specifications.

## 8 Related Work

**Standards landscape analysis.** Baron and Spulber [Baron & Spulber, 2019] provide bibliometric analysis of standards organizations but focus on patents and firm-level strategy rather than technical content. Simmons and Thaler [Simmons, 2019] study IETF participation diversity but do not assess draft content or topical overlap. Our work extends this line by applying NLP techniques to the document content itself.

**AI governance and safety.** Amodei et al. [Amodei et al., 2016] articulate the challenge of aligning AI systems with human values, a concern our safety deficit finding quantifies in the standards context. The EU AI Act [EU, 2024] and NIST AI Risk Management Framework [NIST, 2023] provide regulatory perspectives on AI governance, but neither addresses Internet protocol standardization specifically.

**LLM-assisted evaluation.** Zheng et al. [Zheng et al., 2023] demonstrate that LLM judges can match human evaluation quality for text assessment. Our pipeline extends this approach from evaluating model outputs to evaluating standards documents, using structured prompts for multi-dimensional rating.

**Multi-agent systems.** The AAMAS community has long studied multi-agent coordination [Wooldridge, 2009]. Our analysis reveals that the IETF is now addressing many of the same problems (coordination, trust, resource allocation) but from a protocol standardization perspective rather than an algorithmic one.

## 9 Future Work

1. **Human validation:** Compare LLM ratings against expert assessments for a stratified sample of 30–50 drafts to quantify LLM judge accuracy in this domain.
2. **Longitudinal monitoring:** Deploy the pipeline for continuous analysis as new drafts appear, tracking the evolution of the safety ratio, overlap clusters, and gap coverage over time.

3. **Citation network:** Extract inter-draft references to build a citation graph, enabling influence analysis beyond topical similarity.
4. **Gap-driven standardization:** Use identified gaps to propose new Internet-Drafts—we have already generated five experimental drafts addressing the architectural pillars described in Section 7.4.
5. **Cross-venue analysis:** Extend the methodology to W3C, OASIS, ISO/IEC JTC 1, and 3GPP AI standardization activities for a comprehensive view of the global AI standards landscape.
6. **Multi-LLM validation:** Cross-validate ratings using multiple LLM judges (Claude, GPT-4, Gemini) to assess systematic bias.

## 10 Conclusion

The IETF AI/agent standardization wave represents a unique moment in Internet governance: the community is attempting to standardize the infrastructure for autonomous agents concurrently with their deployment. Our analysis of 434 Internet-Drafts from 557 authors reveals a landscape characterized by both extraordinary energy and significant structural problems.

Three findings demand attention. First, the **4:1 safety deficit**: the community is building agent capabilities four times faster than safety mechanisms, despite the highest-quality proposals being safety-focused. Second, **extreme fragmentation**: 120 competing A2A protocol proposals, 13 independent OAuth-for-agents drafts, and 96% of technical ideas appearing in only one draft indicate that coordination mechanisms are failing to keep pace with submission volume. Third, **organizational concentration**: 18% of all drafts from a single company and approximately 40% from Chinese organizations raise questions about geographic diversity in the standards that will govern global AI agent infrastructure.

The 1,907 technical ideas we extract represent a rich but disorganized design space. The 11 gaps we identify—from behavior verification to human override protocols to cross-protocol translation—highlight where the community’s collective blind spots lie. The architectural vision we sketch, building on existing IETF primitives (WIMSE, ECT, OAuth), suggests a path from fragmentation toward coherence.

The methodology demonstrated here—combining LLM-assisted multi-dimensional rating with embedding-based similarity analysis—is itself a contribution. At \$3.16 in API costs, it provides a scalable, reproducible approach to standards landscape analysis that could be applied to any standards body facing a surge in submissions. As AI standardization accelerates globally, such tools become essential for maintaining coherence and directing limited community attention to the areas that matter most.

The gold rush will not slow down. The question is whether the safety inspectors can catch up.

## Acknowledgments

Analysis was performed using Anthropic Claude (Sonnet 4) for rating, categorization, and idea extraction, and Ollama with nomic-embed-text for embedding generation. We thank the IETF community for maintaining the open Datatracker API that made this analysis possible.

## References

- S. Bradner. The Internet Standards Process – Revision 3. RFC 2026, IETF, October 1996.  
<https://www.rfc-editor.org/rfc/rfc2026>

- J. Arkko. Considerations on Internet Consolidation and the Internet Architecture. RFC 8890 (draft), IETF, 2019.
- J. Simmons and D. Thaler. IETF Participation Trends and Diversity. Presented at IETF 106, 2019.
- J. Baron and D. Spulber. Technology Standards and Standard Setting Organizations: Introduction to the Searle Center Database. *Journal of Economics & Management Strategy*, 27(3):462–503, 2019.
- T. Brown, B. Mann, N. Ryder, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 2020.
- L. Zheng, W.-L. Chiang, Y. Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.
- D. Amodei, C. Olah, J. Steinhardt, et al. Concrete Problems in AI Safety. *arXiv:1606.06565*, 2016.
- M. Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2nd edition, 2009.
- European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, 2024.
- National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, January 2023.
- Google. Agent-to-Agent (A2A) Protocol Specification. Technical report, 2025. <https://github.com/google/A2A>
- Anthropic. Model Context Protocol (MCP) Specification. Technical report, 2025. <https://modelcontextprotocol.io>

## A Composite Score Formula Sensitivity

To verify that our findings are robust to weight choices, we tested three alternative weighting schemes:

Scheme	N	R	M	Mom	O <sup>-1</sup>	Rank corr.
Default	0.30	0.25	0.20	0.15	0.10	1.000
Equal	0.20	0.20	0.20	0.20	0.20	0.96
Maturity-heavy	0.20	0.20	0.30	0.15	0.15	0.95
Novelty-only	0.50	0.20	0.10	0.10	0.10	0.93

Table 10: Spearman rank correlation between composite scores under alternative weighting schemes vs. the default. High correlations ( $\geq 0.93$ ) indicate the rankings are largely robust to weight choice.

## B Keyword Search Terms

Keyword	Rationale
agent	Core term for AI agent drafts
ai-agent	Specific AI agent proposals
llm	Large language model infrastructure
autonomous	Self-operating systems and agents
machine-learning	ML-related protocol work
artificial-intelligence	General AI drafts
mcp	Model Context Protocol ecosystem
agentic	Agentic AI paradigm
inference	AI inference infrastructure
generative	Generative AI protocols
intelligent	Intelligent networking/systems
aipref	AI preference signaling (AIPREF WG)

Table 11: Twelve seed keywords used for Datatracker API queries, with rationale for inclusion.

## C Top Convergent Ideas

Idea	Drafts	Primary Type
Multi-Agent Communication Protocol	8	protocol
Agentic Network Architecture	7	architecture
Cross-Domain Agent Coordination	6	mechanism
ELA Protocol (EDHOC Lightweight Auth)	6	protocol
Agent-to-Agent Communication Paradigm	5	protocol
Action-Based Authorization	5	mechanism
AI Agent Communication Network	5	architecture
Agent Registration Process	5	protocol
AI Gateway	4	architecture
MCP Session Establishment over MOQT	4	protocol
Network Equipment as MCP Servers	4	mechanism
Multi-Agent Interaction Model	4	pattern
Distributed AI Inference Architecture	4	architecture

Table 12: Most frequently occurring convergent ideas (appearing in  $\geq 4$  drafts independently). These represent areas of implicit community consensus.