

The AI-Agent Standardization Surge at the IETF: A Verified Quantitative Survey of 524 Internet-Drafts (2024–2026)

Christian Nennemann
Independent Researcher
ietf@nennemann.de

May 2026

Abstract

Between 2024 and 2026 the Internet Engineering Task Force (IETF) saw a sharp rise in Internet-Drafts addressing AI agents and autonomous systems: monthly submissions in our corpus grew from an average of 3.7 (June 2024–May 2025) to 38.8 (since June 2025), peaking at 106 in March 2026—roughly $35\times$ a typical 2024 month. We present a quantitative survey of this emerging area based on a curated corpus of 524 AI/agent-related IETF Internet-Drafts spanning January 2024 to May 2026. We characterise the corpus along four axes: temporal submission dynamics, thematic category distribution, authorship and working-group structure, and semantic redundancy measured through text embeddings. Two findings stand out: the area is overwhelmingly *pre-standardization*—87% of drafts are individual submissions not adopted by any working group—and it is semantically dense, with 32% of drafts having a near-duplicate (cosine > 0.9) elsewhere in the corpus. Because the thematic categories are produced by an LLM-assisted pipeline, we explicitly quantify their reliability through a two-model re-rating experiment: categorical assignment is substantially reproducible (Cohen’s $\kappa = 0.65$), whereas LLM ordinal *quality* scores such as novelty and overlap are not ($\kappa_w = 0.13$ – 0.21). We therefore report the category landscape but deliberately exclude quality scores from our findings, and we argue this distinction is a general caution for the growing practice of LLM-assisted corpus analysis. All data, queries, and rating artifacts are released for reproduction.

Keywords: IETF, Internet-Drafts, AI agents, standardization, landscape analysis, LLM-assisted classification, inter-rater reliability, text embeddings

1 Introduction

Standardization activity around AI agents and autonomous systems has surged at the Internet Engineering Task Force (IETF). In the corpus studied here, monthly submissions of AI/agent-related Internet-Drafts grew from an average of 3.7 per month over June 2024–May 2025 to 38.8 per month since June 2025, peaking at 106 drafts in March 2026—roughly $35\times$ a typical 2024 month. A surge of this size, concentrated in barely a year, makes the area difficult to navigate: there is no settled taxonomy, the boundaries between proposals are fluid, and it is unclear how much of the activity represents distinct technical work versus overlapping attempts at the same problems.

A natural response is to characterise the space quantitatively, and LLM-assisted corpus analyses are an increasingly common instrument for doing so—using a large language model to classify documents, score them, and summarise trends at a scale that manual reading cannot reach [9, 2]. Such studies, however, rarely report whether the labels they rely on are reproducible. An LLM applied to the same documents under a different model, or even a different

run, may produce different categories and scores; without a reliability check, it is impossible to tell which conclusions rest on stable signal and which on rater noise.

To our knowledge no verified quantitative survey of the AI-agent standardization space at the IETF exists. We address this gap with a descriptive survey that is explicit about which of its measurements can be trusted. This paper is deliberately neutral: it characterises the landscape rather than advocating for any particular proposal or design.

Contributions.

- A curated corpus of **524** AI/agent-related IETF Internet-Drafts spanning January 2024 to May 2026, constructed by keyword candidate selection followed by explicit false-positive filtering (Section 3).
- A characterization of the corpus along four axes—temporal submission dynamics, thematic category distribution, authorship and working-group structure, and semantic redundancy measured through text embeddings (Section 4).
- An explicit two-model inter-rater reliability check that separates the labels we can trust from those we cannot: categorical assignment is substantially reproducible (Cohen’s $\kappa \approx 0.65$), whereas the LLM ordinal *quality* scores are not ($\kappa_w = 0.13$ – 0.21 for the least stable dimensions). We report the former and exclude the latter (Section 5).
- A full release of the data, queries, and rating artifacts used in the analysis, so that every reported number can be recomputed.

Roadmap. Section 3 describes corpus construction, the LLM-assisted classification pipeline, and the embedding setup. Section 4 presents the landscape across the four axes: temporal dynamics (Section 4.1), category distribution (Section 4.2), authorship and working-group structure (Section 4.3), and semantic redundancy (Section 4.4). Section 5 reports the two-model reliability experiment that underwrites the category labels and disqualifies the quality scores. Section 6 discusses implications and limitations, and Section 7 concludes with directions for future work.

2 Related Work

Quantitative analysis of IETF activity. Several studies have examined the IETF corpus empirically. McQuistin et al. characterised the IETF through the lens of RFC deployment, using a dataset of 8,711 RFCs, 4,512 authors, and 2.4 million emails to measure shifts in publication rate, author concentration, and cross-document dependency over time [5]. Zhang et al. extended this line of work with a longitudinal analysis of author affiliations across 2001–2023, covering 73,764 individuals and finding that organisational diversity peaked and then stagnated [13]. The IAB’s own workshop on Analyzing IETF Data (AID, 2021) surveyed open questions about what drives drafts to become RFCs and how community diversity evolves [10]. More recently, Jiménez applied LLMs directly to IETF working-group records to automate the generation of summary reports [3]. Our work complements this body of evidence by providing the first quantitative survey focused *exclusively on the AI/agent topic area*: we document the growth trajectory, authorship structure, and thematic distribution of a cohort that did not exist in large numbers before 2025, and we do so at the Internet-Draft stage rather than the post-publication RFC stage.

LLM-assisted text classification and annotation. The use of LLMs as automated annotators is well established. Gilardi et al. showed that ChatGPT outperforms crowd-workers

on political-science annotation tasks (stance, relevance, frames) across 6,183 tweets, with zero-shot accuracy exceeding MTurk workers by roughly 25 percentage points on average [2]. Tan et al. survey the broader landscape of LLMs for data annotation and synthesis, covering annotation generation, quality assessment, and downstream utilisation [9]. Yang et al. survey emerging AI-agent communication protocols—the class of specification documents our corpus comprises—providing context for the standardisation subjects under analysis [12]. Our contribution in this dimension is not classification per se but the application of the LLM-annotator pipeline to a *technical standards corpus* rather than short social-media texts, and the explicit reliability gate described next.

Inter-rater reliability of LLM labels. The reproducibility of LLM judgments is contested. Reiss identified prompt-sensitivity as a reliability hazard, showing that minor wording changes or repeated identical inputs can shift ChatGPT’s classification output below the thresholds conventionally required for scientific use [8]. Wang et al. found that ChatGPT correlates well with human judgments on several natural-language-generation evaluation dimensions but that the strength of correlation varies substantially across tasks [11]. These results motivate—but do not themselves perform—inter-rater reliability analysis using established statistics. We use Cohen’s κ [1] with the Landis–Koch interpretation scale [4] to report agreement both between two independent LLM re-rating runs and between those runs and the production labels. To our knowledge, no prior quantitative survey of an IETF (or comparable standards-body) corpus has reported this reliability decomposition; the closest precedent is Jiménez [3], who generates LLM summaries of IETF records but does not measure label reproducibility.

Embedding-based semantic analysis of document corpora. We use cosine similarity over dense text embeddings for the redundancy analysis. Reimers and Gurevych established that sentence-level embeddings computed with siamese BERT networks provide efficient and accurate representations for semantic similarity search [7]; more recent work by Nussbaum et al. produces the `nomic-embed-text` model used in our pipeline [6]. The application of embedding-based near-duplicate detection to a *standards corpus* is, to our knowledge, novel: prior IETF studies (e.g., [5]) rely on structural metadata—author lists, citation graphs, working-group membership—rather than semantic similarity over document content.

3 Corpus and Method

3.1 Corpus construction

Internet-Drafts were collected from the IETF Datatracker. Candidate documents were identified by keyword and topic matching for AI-agent and autonomous-system terminology (e.g. “agent”, “autonomous”, “LLM”, “inference”), then filtered to remove false positives—documents in which such terms occur in an unrelated sense (“user agent”, “autonomous system” in the BGP sense). Of 597 candidate IETF documents, 73 (12.2%) were flagged as false positives and excluded, yielding a clean corpus of **524 IETF Internet-Drafts**.

We restrict this survey to IETF Internet-Drafts. The collection pipeline also ingests documents from other standards bodies (ISO, ITU, ETSI, NIST, W3C), but those exhibit heterogeneous and frequently missing publication dates and a different document model; mixing them would compromise the temporal and authorship analyses. All 524 IETF drafts carry ISO-8601 submission timestamps, so the IETF-only scope is also the cleanest with respect to date quality. The corpus snapshot used throughout is `data/drafts.db` as of 2026-05-23.

3.2 Thematic classification

Each draft is assigned to one or more of eleven thematic categories (Table 1) by an LLM-assisted rating pipeline. The classifier is prompted with the draft’s title, metadata, and abstract (truncated to the first 2000 characters), and returns a JSON object containing the category assignment and five ordinal 1–5 quality dimensions (novelty, maturity, overlap, momentum, relevance). We emphasise two properties of this pipeline that bear directly on interpretation:

- **Abstract-only.** Ratings are derived from the abstract, not the full document text. This keeps the pipeline cheap and uniform but means classifications reflect how a draft *presents* itself, not a full reading of its mechanics.
- **Model-derived.** Categories and scores are produced by a large language model (Claude Sonnet, `claude-sonnet-4-20250514`). Section 5 quantifies how reproducible these labels are; the headline result is that we trust the categories but not the ordinal scores.

3.3 Semantic embeddings

For the redundancy analysis (Section 4.4) each draft is embedded into a 768-dimensional vector using the `nomic-embed-text` model run locally via Ollama. Embedding coverage is complete (524/524). We compute cosine similarity over all $\binom{524}{2} = 137,026$ document pairs.

3.4 Reproducibility

The pipeline, queries, and the two-model re-rating used for the reliability analysis are released as scripts (`scripts/survey-phase0.py`, `scripts/rerate-intercoder.py`, `scripts/survey-kappa.py`). The re-rating itself was executed through the Anthropic Batch API at a total cost of USD 2.41. Raw model outputs are released as line-delimited JSON.

4 The Landscape

4.1 Temporal dynamics

Figure 1 shows monthly submission counts for the clean corpus. Activity is negligible through most of 2024 (mean 3.7 drafts/month over June 2024–May 2025) and inflects sharply in late 2025: October 2025 jumps to 33 drafts, and monthly counts since June 2025 average 38.8. The peak month is March 2026 with 106 drafts—roughly $35\times$ a typical 2024 month. We report this as a peak-to-baseline ratio rather than an “average growth rate”: the distribution is dominated by a recent spike, not steady exponential growth. The final two months (April–May 2026) dip relative to the March peak; because drafts continue to be submitted and indexed, the tail of the curve should be read as provisional rather than as a downturn.

4.2 Category distribution

Table 1 gives the distribution of drafts by primary category. The field is concentrated in identity/authentication (141 drafts, 26.9%), agent-to-agent communication protocols (108, 20.6%), and autonomous network operations (64, 12.2%). The sparsest areas are human-agent interaction (5), the residual “Other” bucket (9), and model serving/inference (16); we report these descriptively and draw no normative conclusion about whether sparse areas *ought* to receive more attention. Categories are not mutually exclusive: 92.9% of drafts carry more than one category, reflecting genuine thematic overlap (e.g. a draft may concern both agent identity and authorization). We use the primary (first) category for the distribution.

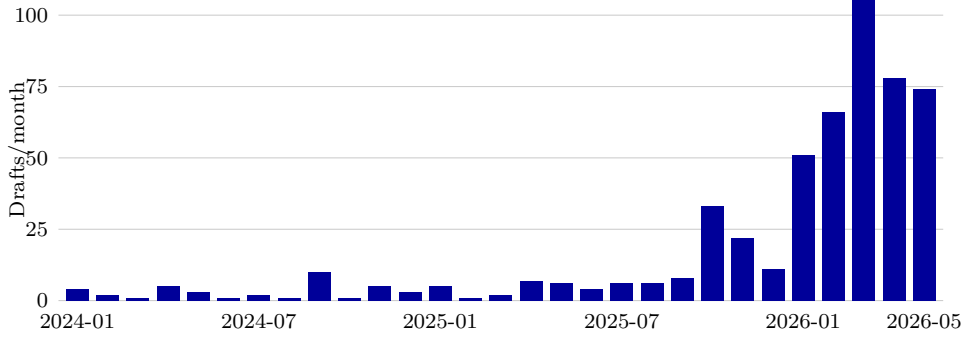


Figure 1: Monthly counts of AI/agent IETF Internet-Drafts, January 2024–May 2026 ($n = 524$). The April–May 2026 tail is provisional (indexing lag).

Primary category	Drafts
Agent identity / authentication	141
Agent-to-agent (A2A) protocols	108
Autonomous network operations	64
ML traffic management	48
Data formats / interoperability	44
Agent discovery / registration	35
Policy / governance	30
AI safety / alignment	24
Model serving / inference	16
Other AI/agent	9
Human-agent interaction	5
Total	524

Table 1: Distribution of the clean IETF corpus by primary category. 92.9% of drafts also carry one or more secondary categories.

4.3 Authorship and working-group structure

The corpus is authored by 619 distinct individuals and is *not* concentrated: the ten most prolific authors together account for only 10.9% of drafts. The most active contributors—Bing Liu (22 drafts), Nan Geng (21), and Zhenbin Li (20)—form a cluster around autonomous network operations.

The strongest structural signal is the maturity of the work. Of the 524 drafts, **456 (87%) are individual submissions** not adopted by any IETF working group; only 28 distinct working groups appear across the remainder. The area is thus overwhelmingly *pre-standardization*: a large volume of individual proposals competing for attention, with comparatively little that has cleared the bar of working-group adoption.

4.4 Semantic redundancy

Pairwise cosine similarity over the 137,026 document pairs has a mean of 0.711 (p90 0.790, p99 0.850, max 1.000). We define a near-duplicate as a pair with cosine > 0.9 : there are 125 such pairs, and **170 drafts (32.4%) have at least one near-duplicate** in the corpus. The four pairs at cosine ≈ 1.0 are legitimate—they are individual versus working-group-adopted versions of the same document (for example `draft-fv-rats-ear` and `draft-ietf-rats-ear`)—rather than data artifacts. Combined with the dominance of individual submissions (Section 4.3), the high redundancy is consistent with a young, crowded design space in which many authors

propose overlapping solutions to the same problems before consolidation occurs.

5 How Reliable Are the Labels?

Every category in Section 4.2 is an LLM judgment. Before treating the distribution as a finding, we ask how reproducible those judgments are. We re-rated all 524 drafts a second time with the same pinned prompt using two models—Claude Sonnet (`claude-sonnet-4-20250514`) and Claude Haiku (`claude-haiku-4-5-20251001`)—and compared the assignments. We report Cohen’s κ for the nominal primary-category assignment and quadratic-weighted κ (κ_w) for the ordinal 1–5 dimensions, interpreting magnitudes by the Landis–Koch convention (< 0.2 slight, 0.2–0.4 fair, 0.4–0.6 moderate, 0.6–0.8 substantial, > 0.8 almost perfect).

Category assignment is substantially reliable. For the primary category, Sonnet and Haiku agree at $\kappa = 0.652$ (raw agreement 70.8%). Each model independently re-rating also agrees substantially with the original production labels (Sonnet $\kappa = 0.645$, Haiku $\kappa = 0.596$). The residual disagreements are not random: they concentrate on semantically adjacent categories (Table 2)—A2A protocols versus autonomous network operations, A2A versus agent discovery, and identity versus the residual “Other” bucket. These are boundary cases, not classifier noise, which is why we treat the distribution as informative while acknowledging \pm a few points of category-boundary uncertainty.

Category A	Category B	Disagreements
A2A protocols	Autonomous netops	17
A2A protocols	Agent discovery/reg	16
A2A protocols	Agent identity/auth	15
Agent identity/auth	Other AI/agent	14
Data formats/interop	Other AI/agent	10

Table 2: Most-confused primary-category pairs between the two re-rating models. Disagreements concentrate on semantic neighbours.

Ordinal quality scores are not reliable. The five 1–5 quality dimensions tell a different and cautionary story (Table 3). Between Sonnet and Haiku, *overlap* reaches only $\kappa_w = 0.127$ (slight—effectively no better than chance) and *novelty* $\kappa_w = 0.206$. *relevance* appears substantial between the two re-rating models (0.728) but collapses to 0.234 against the production labels, indicating it is not stable across runs either. Only *maturity* is consistently moderate (0.59–0.62).

Dimension	κ_w (Sonnet vs. Haiku)	κ_w (Sonnet vs. prod.)
relevance	0.728	0.234
maturity	0.592	0.620
momentum	0.457	0.247
novelty	0.206	0.477
overlap	0.127	0.282

Table 3: Quadratic-weighted κ for the ordinal quality dimensions. Low and inconsistent values—especially for *overlap* and *novelty*—indicate these scores are not reproducible across raters.

Consequence. We therefore report the categorical landscape (Section 4.2) but deliberately exclude the LLM ordinal quality scores from our findings. We regard this split as a general caution rather than an artifact of our particular pipeline: LLMs can place standards documents into a thematic taxonomy with substantial agreement, but asking them to score subjective qualities such as “novelty” or “overlap” on a numeric scale produces labels that do not survive a change of model. Studies that use LLM-assigned quality scores as quantitative evidence should report inter-rater reliability before doing so.

6 Discussion and Limitations

6.1 Implications

Two structural features of the corpus reinforce one another. First, the area is *pre-standardization*: 456 of the 524 drafts (87%) are individual submissions not adopted by any working group (Section 4.3). Second, it is semantically redundant: 170 drafts (32.4%) have at least one near-duplicate (cosine > 0.9) elsewhere in the corpus (Section 4.4). Taken together these point to a young, crowded design space in which many authors independently propose overlapping solutions to the same problems. Such a configuration is consistent with the early phase of a standards effort, where consolidation and competition between proposals have yet to resolve into adopted work; it does not, on its own, indicate duplication of effort in any pejorative sense, since some redundancy is the expected by-product of parallel exploration. We note this dynamic descriptively rather than predicting which proposals will prevail.

Beyond the IETF, our reliability result carries a general caution for LLM-assisted corpus studies. The same pipeline that places documents into a thematic taxonomy with substantial agreement (Cohen’s $\kappa \approx 0.65$) produces ordinal quality scores—“novelty,” “overlap”—that do not survive a change of model (κ_w as low as 0.13). A study that reported the category distribution and the quality scores side by side, without a reliability check, would present a reproducible measurement and rater noise with equal confidence. The discipline of separating the two, by an explicit inter-rater analysis, is what allowed us to keep the former and discard the latter.

6.2 Limitations

Several limitations bound the interpretation of our findings.

- **Abstract-only classification.** Categories and scores are derived from each draft’s title, metadata, and abstract, not its full text (Section 3). Classifications therefore reflect how a draft presents itself rather than a full reading of its mechanics, and a draft whose abstract understates its technical content may be miscategorised.
- **Single snapshot.** The analysis rests on one corpus snapshot (`data/drafts.db` as of 2026-05-23). All counts, trends, and similarities are as of that date and will drift as drafts are added, revised, expired, or adopted.
- **IETF-only scope.** We deliberately restrict the corpus to IETF Internet-Drafts; documents from ISO, ITU, ETSI, NIST, and W3C are excluded because of heterogeneous metadata and a different document model. The survey therefore says nothing about AI-agent standardization outside the IETF, and the overall scale of the field is correspondingly understated.
- **LLM-derived categories.** The thematic taxonomy is produced by a large language model. The two-model κ check (Section 5) mitigates but does not eliminate this dependence: substantial agreement still leaves category-boundary uncertainty of a few points, concentrated on semantically adjacent categories.
- **Provisional recent tail.** Submission counts for the most recent months (April–May 2026) are provisional because of indexing and fetch lag; the dip after the March 2026 peak in

Figure 1 should be read as incomplete data rather than as a downturn.

- **Keyword-based candidate selection.** Candidates were identified by keyword and topic matching, which can both miss relevant drafts whose abstracts avoid the chosen vocabulary and over-include unrelated documents; we observed a 12.2% false-positive rate (73 of 597 candidates) before filtering, and an unknown false-negative rate remains.

7 Conclusion and Future Work

We have presented a verified quantitative survey of AI/agent-related IETF Internet-Drafts, based on a curated corpus of 524 documents spanning January 2024 to May 2026. The corpus exhibits a sharp recent surge—from 3.7 to 38.8 drafts per month, peaking at 106 in March 2026—and is dominated by individual submissions (87% not adopted by any working group) with substantial semantic redundancy (32.4% of drafts having a near-duplicate), the signature of a young, pre-standardization design space. Crucially, we treated the LLM-assisted labels themselves as objects of measurement: a two-model re-rating shows that categorical assignment is substantially reproducible (Cohen’s $\kappa \approx 0.65$), while ordinal quality scores are not ($\kappa_w = 0.13$ – 0.21 for the least stable dimensions), so we report the category landscape and exclude the quality scores.

Several directions extend this work. *Full-text classification* would replace the abstract-only pipeline and test whether categories shift when the classifier reads the complete document. *Longitudinal re-runs* on later snapshots would turn the single-snapshot picture into a moving record of how the surge evolves and whether redundancy consolidates over time. *Cross-SDO extension* to ISO, ITU, ETSI, NIST, and W3C—contingent on reconciling their heterogeneous metadata—would situate the IETF activity within the broader standards landscape. Finally, *tracking working-group adoption* of individual drafts would reveal which of the many competing proposals clear the bar from individual submission to adopted work, giving an empirical handle on how the pre-standardization field resolves.

References

- [1] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- [2] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowdworkers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.
- [3] Jaime Jiménez. Automating IETF insights generation with AI, 2024.
- [4] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- [5] Stephen McQuistin, Mladen Karan, Prashant Khare, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, Waleed Iqbal, Junaid Qadir, and Ignacio Castro. Characterising the IETF through the lens of RFC deployment. In *Proceedings of the 21st ACM Internet Measurement Conference (IMC '21)*, pages 137–149, 2021.
- [6] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.
- [7] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natu-*

ral Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), pages 3982–3992, 2019.

- [8] Michael V. Reiss. Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark, 2023.
- [9] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 930–957, Miami, Florida, USA, 2024.
- [10] Niels ten Oever, Corinne Cath, Mirja Kühlewind, and Colin S. Perkins. Report from the IAB workshop on analyzing IETF data (AID) 2021. RFC 9307, Internet Architecture Board, September 2022. Informational.
- [11] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinnan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? A preliminary study. In *Proceedings of the 4th Workshop on New Frontiers in Summarization (NewSumm@EMNLP 2023)*, 2023. arXiv:2303.04048.
- [12] Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, Weiwen Liu, Ying Wen, Yong Yu, and Weinan Zhang. A survey of AI agent protocols, 2025.
- [13] Yangjun Zhang, Stephen McQuistin, Mladen Karan, Hugo Enrique Ramirez-Centeno, Colin Perkins, Gareth Tyson, and Ignacio Castro. Two decades of IETF affiliations: Evolution and impact. In *Proceedings of the 2025 Applied Networking Research Workshop (ANRW '25)*, 2025.