



Large language models for scientometric mapping of scientific controversy: A validated hybrid AI–Human framework

Teo Susnjak¹ · Cole Palffy¹ · Tatiana Zimina¹ · Nazgul Altynbekova¹ · Kunal Garg² · Leona Gilbert²

Received: 20 February 2026 / Accepted: 30 May 2026 / Published online: 13 June 2026
© The Author(s) 2026

Abstract

This study advanced large language model (LLM)-enabled scientometric methods for analysing contested scholarly communication through a hybrid AI–human workflow for large-scale mapping of stance and themes in scientific corpora. We treated the literature as a large textual record and examined how competing knowledge claims were distributed across venues and over time. We used the Lyme disease controversy as a testbed, applied multiple LLMs to thousands of abstracts, and validated the outputs with domain experts to trace longitudinal shifts in authorial stance and thematic focus. We examined how alternative perspectives were distributed across venues and themes, how these distributions changed over time, and how venue stratification shaped the visibility of different epistemic positions. Results showed persistent asymmetries in publication placement and citation attention, with higher-impact venues disproportionately hosting one perspective relative to the other, while thematic mapping identified recurring fault lines around disputed claims. Validation showed substantial alignment between LLM outputs and expert ratings across classification tasks (Cohen’s kappa 0.58–0.71), broadly comparable to expert–expert agreement and sufficient for corpus-level scientometric analysis. The framework offers a reproducible model of human–LLM integration in literature-based scientometrics and enables venue- and time-resolved indicators for monitoring controversies and publication imbalances.

Keywords Bibliometric big data · Information structures · Computational discourse analysis · Knowledge mapping · Scientometrics · Scholarly information ecosystems · Large language models · Stance detection · Information infrastructures · Scholarly communication · Scientific controversy · Lyme disease controversy

✉ Teo Susnjak
t.susnjak@massey.ac.nz

¹ School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

² Tezted Oy, Jyväskylä, Finland

Introduction

Scientific literature now grows too quickly for manual analysis, especially in contested areas such as healthcare (Bornmann et al., 2021). Digital publication archives record not only findings but also epistemic commitments, institutional alignments, and scientific controversies (Boyd and Crawford, 2012). Manual thematic coding in these contexts remains valuable, but it does not scale to the current volume of literature (Borah et al., 2017; Mu and Wang, 2022). Conversely, standard bibliometric indicators or basic topic modelling often fail to capture subtle stance-taking, framing contests, and shifting epistemic positions within scientific debates (Hoyle et al., 2021). These limits create a clear need for validated, scalable methods for literature-based scientometric analysis.

Recent advances in large language models (LLMs) (Brown et al., 2020) create new opportunities for semantic and stance-aware analysis of scientific texts. Yet LLM-assisted scientometrics still lacks well-validated workflows and structured evaluation protocols that integrate automated outputs with human oversight (Thelwall, 2026; Bharti et al., 2026; Murray et al., 2025).

We close this gap with a hybrid AI–human framework for large-scale, longitudinal analysis of contested scientific discourse. We treat stance and theme (or topic) labels as measurement outputs, then aggregate them into time-, venue-, and citation-stratified indicators that can be evaluated for reliability and used for comparative scientometric analysis. We demonstrate it on the Lyme disease controversy, which is a highly polarised, multi-decade debate over persistent symptoms after standard antibiotic treatment (Elliott, 2021). Within our design, domain experts remain essential: they define the classification schema, refine boundary cases, and validate outputs, while the LLM applies that expert-designed schema consistently across a corpus too large for manual coding alone. Lyme disease is a tick-borne bacterial infection caused by the *Borrelia burgdorferi* group and transmitted by *Ixodes* ticks. Most cases resolve after 2–4 weeks of antibiotics (Wormser et al., 2006). The controversy centres on the *cause* of ongoing symptoms in a subset of patients. The main-stream view (Post-Treatment Lyme Disease Syndrome, PTLDS) attributes them to immune dysregulation; the opposing view (Chronic Lyme Disease, CLD) posits persistent infection (Wormser et al., 2006; Cameron et al., 2014). These positions have generated a substantial, divided literature (Tonks, 2007; Dumes, 2020), making Lyme a strong demonstration case for stance detection and indicator construction.

This study contributes a validated human–LLM framework for literature-based scientometric analysis. We combine frontier¹ LLMs with structured expert validation to detect stance (Küçük and Can, 2020; Lan et al., 2024) and analyse thematic change across thousands of scholarly abstracts spanning 25 years (2000–2024). Rather than relying on keyword counts or unsupervised topic models, the framework uses explicit classification schemas, iterative refinement, and inter-rater benchmarking against domain experts. It yields reproducible time- and venue-resolved indicators derived from LLM-labelled corpora. We organise the study around four research questions:

¹ The LLMs used at the time of the experiments were state-of-the-art, acknowledging that ever more capable models continue to be released at pace.

- **RQ1.** From a methodological validation perspective, to what extent can a hybrid human–LLM framework reliably reproduce expert judgement in complex stance classification tasks within contested scientific domains?
- **RQ2.** With a focus on longitudinal change, how have competing perspectives in the Lyme disease literature shifted over time?
- **RQ3.** In considering venue-level imbalances, how do publication venues differentially host and amplify competing perspectives when aggregated into LLM-derived scientometric indicators?
- **RQ4.** How are thematic patterns and explanatory models distributed across the competing perspectives in this controversy?

While informed by Science and Technology Studies (STS) concepts regarding controversy and knowledge construction (Latour and Woolgar, 2013; Cetina, 1999), our primary objective was methodological, namely, to demonstrate a transparent, reproducible, and validated framework for LLM-empowered scientometrics. The Lyme corpus served as an illustrative and information-rich testbed with relatively legible stance categories, allowing us to evaluate the robustness, limits, and indicator potential of the proposed framework. We do not assume that all controversies will be equally well structured; cases with weaker boundaries, more than two salient positions, or greater conceptual drift would require adapted schemas and potentially more flexible classification designs.

Related Work

Limitations of traditional methods for scientific literature analysis

Researchers have employed diverse computational techniques for biomedical and other literature analyses. Traditional topic modelling methods, such as Latent Dirichlet Allocation (LDA) have been used most frequently to uncover latent thematic structures in scientific corpora (Xie and Waltman, 2025). However, such topic modelling approaches frequently rely on simplistic preprocessing that may retain irrelevant contextual material of a given study, such as background and limitations, thereby potentially introducing noise that degrades topic quality (Zhang and Chen, 2025). These approaches are also limited in explicitly capturing author stance or evaluative framing in contested domains (Zhang et al., 2020). Embedding-based topic models combine contextualised document embeddings with dimensionality reduction and clustering to induce semantic groupings (e.g., BERTopic) (Grootendorst, 2022). However, their outputs can be harder to interpret than LDA (Xie and Waltman, 2025), and they still require expert judgement to validate and label topics. In contested domains, stance is therefore better treated as a distinct modelling target rather than inferred from topical structure alone. On the other hand, sentiment analysis is widely applied in social media and review settings, but its effectiveness is highly domain-dependent and sensitive to linguistic structure, including negation and contextual ambiguity (Yousif et al., 2019). These challenges become particularly salient in formal scholarly writing, where authors often introduce claims and external viewpoints using stance-neutral formulations and defer evaluation across larger stretches of discourse rather than encoding it lexically (Sawaki, 2014). As a result, surface sentiment cues can be sparse or misleading, degrading the suitability of sentiment-based methods for literature-based scientometric interpretation in polarised domains. Bibliometric techniques such as

co-authorship networks and citation analyses show collaboration structure, influence, and broad field-level patterns, but they rely on bibliographic metadata and do not capture the rhetorical or stance-bearing content that drives controversies (Ullah et al., 2022). Similarly, information extraction methods (e.g., NER and relation extraction) used commonly in the biomedical domain, though effective at capturing entities and factual relationships, tend to overlook implicit stance and rhetorical framing that are crucial in contentious scientific debates (Lee et al., 2020). Meanwhile, manual coding, while valuable for deep interpretive work, becomes unfeasible at scale, particularly given the time and labour required to ensure consistency and inter-rater reliability across large corpora (Borah et al., 2017). Recent scientometric work has moved toward altmetrics, social-media signals, and post-publication platforms (Costas et al., 2015; Haustein et al., 2014; Ortega, 2022) which we seek to extend further by adding semantic stance and theme classification inside the literature itself, turning LLM labels directly into venue- and citation-based indicators.

Advancements in stance detection for scientific discourse

Stance detection is commonly formalised as a supervised classification problem in which a model predicts a writer's position expressed in a text with respect to a specified target, typically using label sets such as *Favor*, *Against*, *None*, and sometimes adding a *Neutral* label, with ongoing debate about whether *Neutral* is meaningfully distinct from *None* when a text contains no stance signal toward the target (Alturayef et al., 2023; Khiabani and Zubiaga, 2025). A closely related framing treats stance as the relation between two texts, which is especially prevalent in misinformation and credibility pipelines, for example when estimating whether a headline aligns with an article body, or whether an evidence passage supports or contradicts a claim (Gera and Neal, 2025; Alturayef et al., 2023).

Methodologically, stance detection has moved from feature-engineered classifiers to transformer encoders and, more recently, to prompt-based LLM inference in zero- and few-shot settings (Gera and Neal, 2025; Khiabani and Zubiaga, 2025). In scientific discourse, stance is rarely signalled by stable lexical markers. It must instead be inferred from how a claim is positioned relative to prior work, evidence strength, and domain conventions. Newer models improve general language understanding, but they may still degrade under domain shift because they learn decision rules tied to the training distribution's terminology, topics, and rhetorical patterns (Gera and Neal, 2025; Khiabani and Zubiaga, 2025). Recent literature also emphasises knowledge-enhanced prompting, where external resources such as lexicons, knowledge graphs, or retrieved context improve generalisation, alongside strategies such as chain-of-thought and self-consistency for resolving implicit stance (Gera and Neal, 2025; Khiabani and Zubiaga, 2025).

Language models in biomedical language processing

Domain-adapted transformers such as BioBERT, PubMedBERT, and SciBERT have improved biomedical text processing by adding domain-specific context and semantics (Lee et al., 2020). Recent general-purpose LLMs also perform well on biomedical reasoning, literature analysis, and classification (Arhiliuc et al., 2025; Singhal et al., 2025). At the same time, their use in biomedical scientometrics raises familiar concerns about hallucination, prompt sensitivity, and model opacity (Li et al., 2025a). Recent techniques, including self-reflection prompting, enable LLMs to critique and refine their outputs, improving consistency for structured classification tasks such as stance detection (Renze and Guven,

2024), while prompt optimisation frameworks have demonstrated promise to address prompt fragility (Li et al., 2025a).

Hybrid AI–Human approaches in scholarly text analysis

In recognition of the limitations of fully automated approaches, hybrid AI–human methodologies have become more prominent in biomedical natural language processing (NLP), particularly for complex literature analysis tasks. These approaches integrate human expertise via active learning, expert validation, and iterative refinement, improving accuracy and auditability (Gilardi et al., 2023). Semi-automated screening, annotation, data extraction, and stance classification increasingly leverage AI for initial labelling, followed by expert oversight to confirm or correct interpretations, balancing computational scalability with methodological rigour (Li et al., 2025b). In scientometrics, such hybrid designs are also relevant to concerns about low-quality or fabricated content entering the scholarly record and the need for reliable validation practices (Cabanac and Labbe, 2021; Dadkhah et al., 2023). Beyond immediate quality control, hybrid workflows also foreground a broader methodological issue, namely, that LLM-based indicators are not neutral outputs but constructed measurement instruments that require explicit design and evaluation. More recent scientometric work underscores this point by arguing that language models are already being incorporated into bibliometric workflows, but their rapid uptake has outpaced corresponding theoretical and empirical foundations, motivating explicit workflow design and evaluation frameworks (Murray et al., 2025). Complementary guidance for responsible research evaluation highlights transparency, verifiability, and the continued role of expert judgement when LLM-derived measures are used for evaluative purposes (Thelwall, 2026).

Lyme disease as a testbed for literature-based scientometrics

Sociological analyses highlight Lyme disease as epistemically contentious, shaped by competing evidence domains as well as alternative information ecosystems, institutional dynamics, and public discourse that contest prevailing biomedical framings (Puppo et al., 2023; Olechnowicz et al., 2023). Lyme disease has consequently become a paradigmatic biomedical controversy, defined by polarised perspectives on the causes and treatment of persistent symptoms following standard antibiotic therapy. The mainstream position, endorsed by bodies such as the IDSA², attributes these symptoms to immune or inflammatory after-effects, now labelled PTLDS (Wormser et al., 2006). In contrast, proponents of CLD assert that persistent infection underlies these symptoms and advocate extended antibiotic regimens (Cameron et al., 2014; Rebman et al., 2017). These conflicting perspectives have produced a divided research landscape shaped by scientific, institutional, and patient-based epistemologies (Tonks, 2007; Peretti-Watel et al., 2019).

As shown in Figure 1, the controversy has unfolded across several decades. From the emergence of Lyme disease in the 1970s to the formation of advocacy movements and intensified media framing in the 2000s, the discourse reflects tensions between clinical authority and patient activism (Elliott, 2021; Stricker and Johnson, 2008; Lantos, 2011). More recently, attention has shifted toward diagnostic ambiguity, with increasing emphasis on patient narratives and the role of online platforms in shaping public understanding. For

² Infectious Diseases Society of America

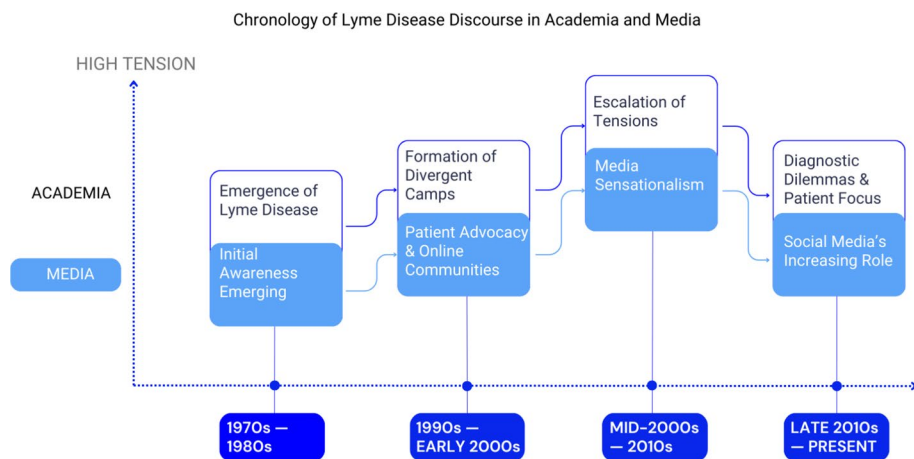


Fig. 1 Key phases and characteristics of the Lyme disease discourse, illustrating the evolution of themes and polarisation relevant to computational analysis of the corresponding literature (Elliott, 2021; Stricker and Johnson, 2008; Lantos, 2011)

our purposes, this long-running, polarised discourse yields an information-rich corpus that is well suited to evaluating LLM-based stance and thematic classification, and to demonstrating how such labels can be aggregated into time- and venue-resolved indicators for contested scholarly communication. We therefore treat Lyme disease as a strong methodological demonstration case rather than as a universal proxy for all scientific controversies. Its controversy is unusually well documented and its stance categories are comparatively legible, which helps with workflow evaluation but also bounds direct generalisation to less clearly structured cases. Our focus is therefore methodological, using the Lyme controversy to evaluate a literature-based scientometric workflow rather than to assess clinical efficacy.

Methodology

Our workflow integrates LLMs with structured expert validation and draws on STS perspectives on scientific controversy. It has four stages: (1) data acquisition, (2) automated classification using LLMs, (3) validation through expert review, and (4) thematic analysis with human-in-the-loop verification. Figure 2 summarises the workflow, and the subsections below describe each stage. All data not subject to copyright restrictions, prompts, and outputs are openly available³.

³ GitHub repository <https://github.com/teosusnjak/Lyme-disease-controversy>

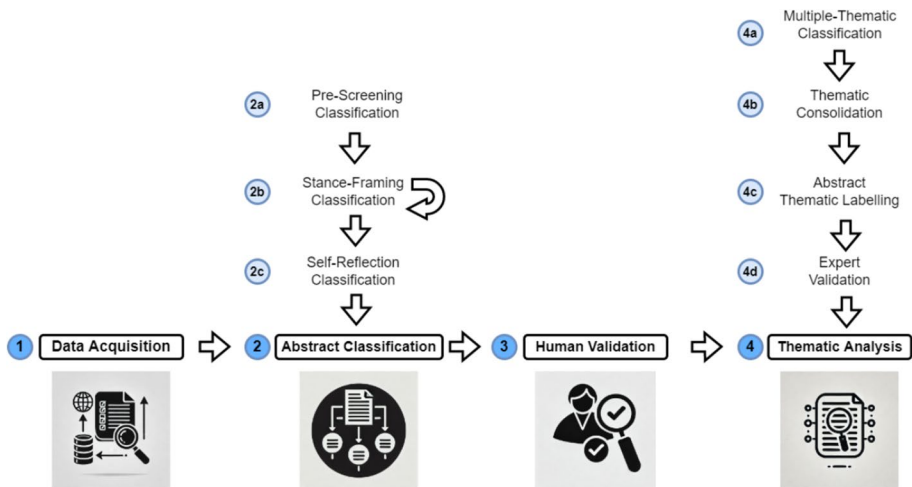


Fig. 2 Overview of the steps comprising the proposed hybrid AI-driven content analysis methodology

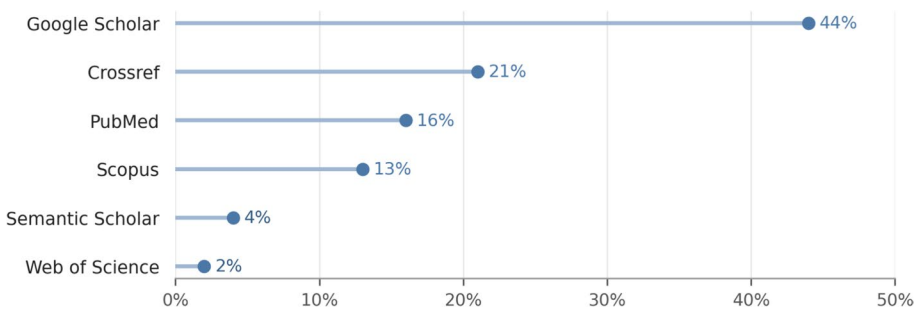


Fig. 3 Proportion of retrieved articles per database

Dataset acquisition - Step 1

Step 1 in Figure 2 shows the data acquisition process. We collected records for 2000–2024 using the Publish or Perish (PoP) software⁴ across academic databases. Figure 3 shows the share of abstracts retrieved from each source. We searched paper titles for *lyme* and matched documents using combinations of *disease*, *borreliosis*, *borrelia*, *chronic*, *controversy*, *post-treatment*, *PTLDS*, *acute*, *syndrome*, and *post-Lyme*, which yielded an initial dataset of 84,140 studies.

Data screening and filtering

Next, the PRISMA flow diagram in Figure 4 details the pre-processing, screening and filtering process. Records with missing or duplicated DOIs were excluded. More than 38,000

⁴ This data collection builds on and extends the earlier work (Susnjak, 2024).

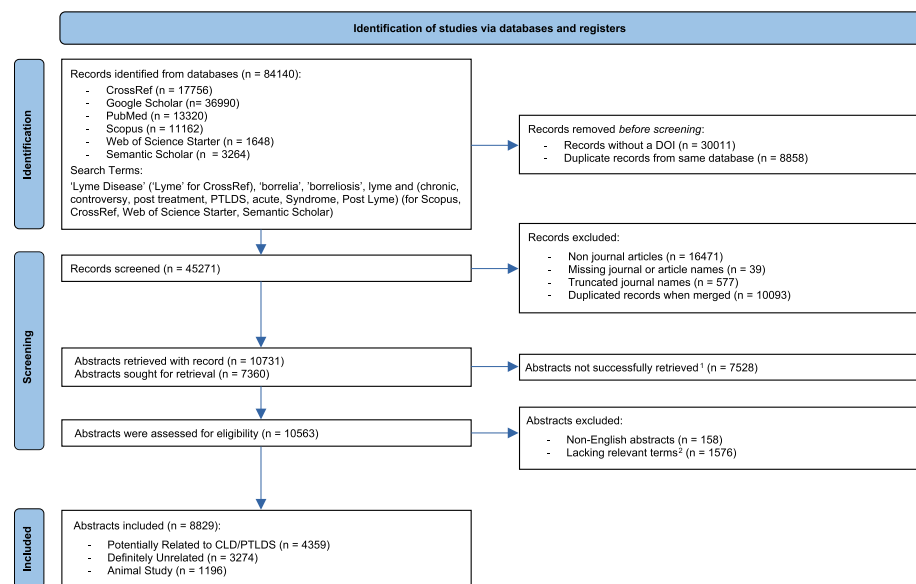


Fig. 4 PRISMA flow diagram detailing the corpus construction and filtering process for the Lyme disease dataset. ¹This number includes 2088 abstracts with less than 300 characters in length, which were rejected due to the information content being too low for analysis. ²Lyme, Borrelia*, burgdorferi, Ixodes, Erythema, migrans, tick-borne, tickborne, tick borne

records were removed, leaving 45,271 records for screening. Subsequently, 27,180 records were excluded due to missing publication names, article titles, or abstracts. Python scripts were written to automatically retrieve missing abstracts where possible using APIs and DOIs⁵, resulting in 7,360 previously missing abstracts. 7,528 records were excluded due to irretrievable abstracts, non-English text, and a lack of relevant search terms. Some abstracts remained irretrievable, and because these exclusions may not be random, reduced corpus completeness may affect observed stance prevalence. The resulting dataset should therefore be interpreted as a large, multi-database record of the accessible English-language literature rather than a complete census of all potentially relevant publications. Following this step, the resulting dataset comprised 8,829 potentially relevant abstracts, requiring more detailed screening and analysis for relevance.

Abstract classification - Step 2

After rule-based screening and filtering, 8,829 records remained for semantic assessment of relevance to the PTLDS/CLD controversy. We then applied three increasingly specific classification steps. Step 2a used a general-purpose LLM (OpenAI GPT-4o-mini, temperature = 0.0) to remove clearly irrelevant abstracts with high confidence while retaining uncertain cases for deeper analysis. Each abstract was classified into one of three predefined classes: *Potentially Related to CLD/PTLDS*, *Definitely Unrelated*, or *Animal Study*.

⁵ Scopus API; Springer Nature API

Prior work supports the use of machine learning for medical text classification (Cohen et al., 2006), and recent LLM-based workflows have shown competitive performance on related screening and extraction tasks (Li et al., 2025c; Gilardi et al., 2023).

We processed each abstract individually via API calls using the classification prompt (Appendix A). The model returned a JSON object containing the abstract index, a classification, and a confidence score (*High*, *Medium*, or *Low*). Abstracts classified as *Potentially Related to CLD/PTLDS* were retained for further analysis. Low-confidence classifications outside this category were flagged for additional validation to reduce false exclusions. This pre-screening step categorised the abstracts as follows: 4,160 (48.2%) *Potentially Related to CLD/PTLDS*, 3,274 (37.9%) *Definitely Unrelated*, and 1,196 (13.9%) *Animal Study*. All *Animal Study* and *Definitely Unrelated* abstracts receiving Medium to High confidence were eliminated from further analysis.

Stance-Framing classification - Step 2b

In Step 2b, we classified the remaining 4,160 abstracts into categories aligned with the study's aims. This required the model to identify both implicit and explicit positions within a contested domain.

From stance to frame detection

Traditional sentiment analysis relies on lexicon-based polarity counts and often struggles with the neutrality and domain-specific terminology common in academic texts (e.g., “disease” flagged as negative) (Wankhade et al., 2022). Stance detection, which classifies text as supporting, opposing, or neutral toward a target (Küçük and Can, 2020), is better suited to this task, and recent LLMs perform well on it (Lan et al., 2024). In Step 2b, we broadened “stance” to capture interpretive *frames*: the implicit perspectives, assumptions, and reasoning styles authors adopt in the CLD/PTLDS debate (Entman, 1993; Fairclough, 1992). A PTLDS frame, for example, may emphasise immune-mediated explanations, whereas a CLD frame may foreground persistent infection. This extension lets the model identify deeper semantic cues within the corpus. We engaged subject experts in an iterative refinement process (Step 2b, Figure 2) to design effective classification prompts. The final prompt specified definitions, classification criteria relevant to known points of contention, and few-shot examples, directing the LLM to label 4,160 abstracts as “Supports PTLDS”, “Supports CLD”, “Neutral”, “Unrelated”, or “Animal Study”. The LLM was also tasked with providing a justification text and a confidence score for each classification. This process yielded 1,033 abstracts classified as “Supports PTLDS”, “Supports CLD” or “Neutral”. The prompts and example outputs are detailed in the Appendix (Figures 11 and 12).

Self-reflection classification - Step 2c

The classifications and justifications from Step 2b were then re-evaluated using automated methods and corrected where necessary to improve consistency. To achieve this, we used a self-reflection prompting technique (Renze and Guven, 2024), in which the model reassesses and refines its initial outputs to identify and correct errors. After executing this step, 5.3% (229 out of 4,359) of classifications changed. The largest category to undergo reclassification was “Neutral” (98), where 60% flipped to “Supports PTLDS” upon revision, and “Unrelated” (64), where 89% subsequently became “Neutral”. Only two abstracts

underwent a major revision from “Supports CLD” to “Supports PTLDS”, supporting the stability of the approach. Several examples of classifications and adjustments are again provided in the Appendix B.2 for transparency.

Human validation (Step 3)

To assess whether automated stance classifications are reliable and aligned with expert judgement, we incorporated a structured human validation process. In this phase (Figure 2), outputs from the primary LLM classifier (including the self-reflection refinement) were compared with annotations from two domain-expert raters and with classifications obtained from additional state-of-the-art LLMs, enabling assessment of consistency and replicability. The two human raters were co-authors and biomedical scientists with domain expertise and familiarity with the Lyme disease and chronic illness literature relevant to this study. In this study, GPT-4o-mini was the primary LLM used for classification; six additional LLMs were included in validation: Google Gemini 2.0 Flash (and Gemini Thinking mode), Anthropic Claude 3.5 Sonnet, DeepSeek R1, Alibaba Qwen 2.5 Max, and X Grok-3.

Inter-rater reliability (IRR) analysis (Cohen, 1960; Fleiss, 1971) was used to compare agreement between: (1) the primary LLM and the two expert raters, (2) the two human raters, and (3) all annotators, including the additional LLMs. IRR was quantified using Cohen’s Kappa (Cohen, 1960) for pairwise comparisons and Fleiss’ Kappa (Fleiss, 1971) for agreement among multiple raters, where Cohen’s Kappa is defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad (1)$$

where P_o represents the observed agreement and P_e is the expected agreement by chance. Standard interpretation thresholds suggest that $\kappa < 0.40$ corresponds to poor agreement, values in the range 0.40–0.60 indicate moderate agreement, 0.60–0.80 denote substantial agreement, and $\kappa \geq 0.80$ reflects almost perfect agreement (Mchugh, 2012; Sim and Wright, 2005). Fleiss’ Kappa generalises this measure for multiple raters, providing a single overall metric that accounts for chance agreement across all annotators.

The primary purpose of this validation step (reported in the Results Section [Method Performance and Validation \(RQ1\)](#)) was to establish that the LLM-based classifications, augmented with self-reflection prompting, exhibit reliability comparable to that of human experts. In this context, robustness refers to measurement stability: if the LLM agrees with expert raters at a level comparable to expert–expert agreement, then the coding framework is being applied consistently enough to support corpus-level scientometric analysis. This does not establish an objective ground truth for every individual label, but it does indicate that the resulting classifications are not merely idiosyncratic to a single rater or model.

Prior to annotation, both human raters underwent training to ensure consistency in applying the classification framework (see stance-framing prompt in the Appendix A). This entailed meetings and the provision of a classification training pack. The raters were provided with the same prompts, criteria, and assumptions used by the LLMs, along with classification examples for reference. A custom web application was developed for independent annotation (Appendix Figures 12 and 11), enabling raters to classify a random sample of 150 abstracts. For each abstract, raters selected a classification and chose between two possible justification options for their decision. These justification options were visible within the annotation interface during the entire task, and raters were not given the option to reject

both and enter free-text alternatives. We used this bounded design to keep the validation workflow structured and analytically tractable, but acknowledge it also creates a possible source of anchoring bias that should be considered when interpreting the justification-agreement results. For interpretive analysis of disagreement patterns in Section [Method Performance and Validation \(RQ1\)](#), we additionally report a stance-focused subset consisting of the 90 abstracts in this validation sample whose classifications were “Supports PTLDS”, “Supports CLD”, or “Neutral”.

Thematic analysis - step 4

Identification and classification of overarching themes

After validating the stance labels for 1,033 abstracts (*Supports PTLDS*, *Supports CLD*, *Neutral*), we conducted a hybrid thematic analysis to identify recurring reasoning patterns. As illustrated in Figure 2, the process combined: (1) LLM-driven extraction of themes from abstracts and justification texts; (2) cross-model reconciliation to align independently generated theme sets; and (3) expert review for conceptual validation. This workflow aimed to balance LLM scalability with targeted human interpretation.

Methodological approach description

The thematic identification process proceeded in four structured phases, aiming to ensure methodological rigour and conceptual validity, drawing upon established qualitative thematic analysis principles (Braun and Clarke, 2006; Schreier, 2012):

1. Multiple-Thematic Identification (Step 4a):

- A textual dataset was compiled comprising justification texts from Steps 2b and 2c. Due to LLM context-window limitations, the dataset was reduced to a random sample of 800 (out of 1,033) justification texts. This was the maximum feasible size for effective processing while providing a substantial corpus for thematic exploration.
- Multiple high-capability reasoning LLMs were used for theme identification to reduce dependence on any single model. Three models—GPT preview-01, Gemini 2.0 Flash, and DeepSeek R1—were each tasked with identifying overarching themes.
- Each LLM was instructed to identify themes without a predefined number, and to propose “half a dozen or more overarching themes” by clustering recurrent arguments and semantic patterns.
- This process produced three independently generated theme sets, requiring reconciliation.

2. Thematic Consolidation (Step 4b):

- Despite operating independently, all three models converged on eight overarching themes (Table 1), suggesting consistency in the underlying thematic structure and reducing reliance on any single model’s idiosyncrasies.
- To establish face validity, two subject-expert co-authors reviewed the LLM-generated themes against the Lyme disease literature and relevant STS concepts, confirming that each theme was distinct and conceptually coherent.

Table 1 Summary of themes independently derived from three LLMs and their mapping. The appendix offers a deeper explanation of the themes and their significance.

GPT preview-o1	Gemini 2.0	DeepSeek R1
Persistence vs. resolution of infection	Persistence vs. Resolution of Infection	Etiological Mechanisms: Persistent Infection vs. Post-Infectious Immune Responses
Diagnostic uncertainty and misdiagnosis	Diagnostic Uncertainty and Misdiagnosis	Diagnostic Complexity and Biomarker Development
Effectiveness of antibiotic therapy	Effectiveness of Antibiotic Therapy	Clinical Management Controversies
Role of immune dysregulation	Immune Dysregulation	Autoimmune Pathways and Residual Antigenic Debris
Psychological vs. biological basis	Neurocognitive and Neuropsychiatric Manifestations	Long-Term Outcomes and Symptom Heterogeneity
Subjectivity vs. objectivity of symptoms	Patient-Centred Experiences	Advocacy and Psychosocial Burden
Sociocultural and ethical factors	Sociocultural and Ethical Factors	Sociocultural and Institutional Influences
Mechanisms of pathogen persistence	Mechanisms of Pathogen Persistence	Bacterial Pathogenesis and Host Interactions

- For reconciliation, we aligned themes from all three LLMs (Table 1) using an inductive–deductive process. First, GPT-`preview-01` synthesised a consolidated theme set; then the same experts reviewed and refined this output—removing redundancies, clarifying scope, and ensuring consistency with prior conceptual framing (Kuhn, 1962).
 - Finally, we consolidated the eight themes into a unified framework (definitions in Supplementary Material). This framework combines LLM-driven pattern recognition with expert interpretation to support conceptual coherence in thematic analysis (Braun and Clarke, 2006; Schreier, 2012).
3. Abstract Thematic Labelling (Step 4c): Each of the 1,033 abstract–justification pairs was automatically assigned up to two themes by GPT-`4o-mini`, acknowledging the discourse’s multidimensional nature.
 4. Expert Validation (Step 4d): We randomly sampled 50 abstract–justification pairs for expert review; each pair was evaluated by two validators, yielding 100 validation decisions in total. Validators confirmed 96% of the labels, supporting the thematic classification approach.

Linking classifications to scientometric indicators

To translate abstract-level classifications into scientometric indicators, we aggregated stance labels by publication year and venue. We calculated a descriptive journal stance-gap score by computing the percentage-point difference between PTLDS-aligned and CLD-aligned articles published in each venue. To benchmark journals against the overall corpus while accounting for temporal change, we additionally calculated time-adjusted observed-minus-expected stance shares, where expected values were derived from the year-specific corpus distribution weighted by each journal’s publication-year mix. Citation impact was analysed by cross-referencing stance labels with citation counts retrieved from Scopus, allowing us to quantify the epistemic influence of specific framing strategies.

Reproducibility and data availability

To ensure transparency and reproducibility, the filtered corpus of abstract details (excluding abstract text due to copyright obligations), prompt templates, and classification outputs are openly available via our GitHub repository⁶.

Results

Method performance and validation (RQ1)

We begin by presenting the IRR results of the Step 2b classification task, in which abstracts were annotated using the full operational label set of “Supports PTLDS”, “Supports CLD”,

⁶ GitHub repository <https://github.com/teosusnjak/Lyme-disease-controversy>

“Neutral”, “Unrelated”, and “Animal Study”, in order to validate the proposed methodology (**RQ1**). Before considering agreement between LLMs and human raters, we establish a baseline IRR by assessing agreement between two domain-expert human raters from the author team. This yielded a Cohen’s Kappa ($\kappa = 0.501$), indicating moderate agreement and aligning with established IRR benchmarks (Mchugh, 2012; Sim and Wright, 2005) and prior studies of subjective classification tasks such as qualitative content analysis, medical diagnosis, and thematic coding (Hallgren, 2012). Perfect agreement is rarely expected in complex classification tasks due to interpretive differences and variation in emphasis on textual elements.

Next, we evaluated agreement between the final revised classifications (produced via GPT-4o-mini’s self-reflection process, referred to as “GPT” in Table 2) and alternative LLM-generated classifications (Gemini 2.0 Flash, Gemini 2.0 Flash-Thinking, Claude, DeepSeek, Qwen, Grok-3), alongside the original GPT classifications and those of the two human raters. Table 2 presents Cohen’s Kappa for each pairwise comparison.

Key observations emerge from Table 2:

- The strongest agreement occurred between the original and revised GPT classifications ($\kappa = 0.767$), indicating that self-reflection refined classifications without fundamentally altering them. This suggests that initial classifications were internally consistent and required only minor adjustments, supporting the stability of the prompt and labelling scheme.
- The agreement between GPT’s revised classifications and the human raters ($\kappa = 0.583$ for Interrater 1; $\kappa = 0.508$ for Interrater 2) is comparable to the human–human agreement ($\kappa = 0.501$). This indicates that the LLM’s revised classifications are at least as consistent with each human rater as the human raters are with each other, supporting the validity of the automated labels for this task. The ambiguity of the literature and the contested nature of the domain plausibly contribute to the moderate κ values, rather than indicating rater failure. Additionally, on the subset of validation cases where both human raters agreed ($n = 91$), agreement with GPT’s classification was higher ($\kappa = 0.709$). With respect to model choice, the highest agreement with both human raters occurred for GPT and the original GPT classifications, supporting the suitability of the chosen model for this classification task.

Table 2 Cohen’s Kappa agreement between GPT revised classifications after self-reflection and all other raters/classifications

Comparison	Cohen’s Kappa
GPT vs. Original GPT classification	0.767
GPT vs. Gemini	0.717
GPT vs. Gemini-Thinking	0.608
GPT vs. Qwen	0.600
GPT vs. Claude	0.592
GPT vs. Human interrater 1	0.583
GPT vs. Human interrater 2	0.508
GPT vs. DeepSeek	0.475
GPT vs. Grok-3	0.458
Human–Human agreement	0.501

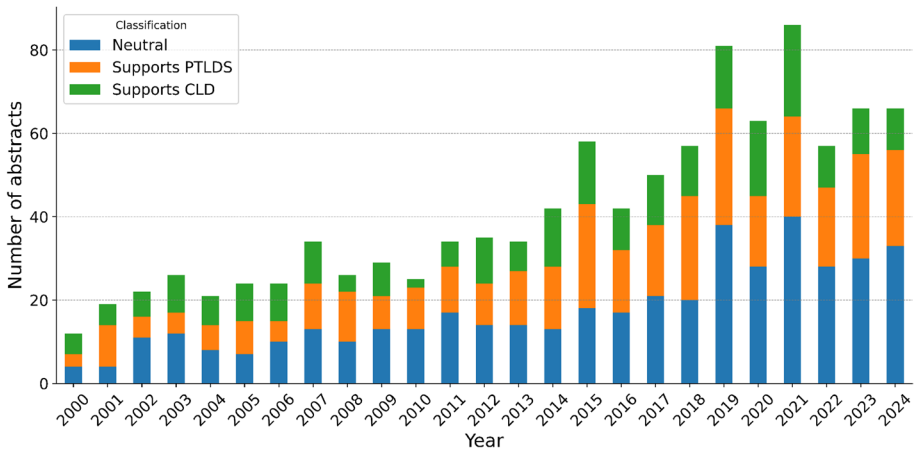


Fig. 5 Number of relevant studies by year (2000–2024) and their classification. Classifications consisted of Neutral (blue), Supports PTLDS (orange), and Supports CLD (green)

- The highest agreement with an alternative model occurred with Gemini ($\kappa = 0.717$), followed by Qwen ($\kappa = 0.600$), indicating broadly similar classification patterns to GPT’s revised outputs.
- The lowest agreement was observed with Grok-3 ($\kappa = 0.458$) and DeepSeek ($\kappa = 0.475$), indicating notable divergences in their outputs.

For completeness, to assess global agreement among all human and automated classifiers, we computed Fleiss’ Kappa ($\kappa = 0.537$). This indicates moderate agreement, aligning with the baseline human–human agreement and reinforcing consistency of the classification framework as a whole. Fleiss’ Kappa further supports the LLM-assisted methodology by showing that the aggregate classification structure remains coherent despite variability across models, consistent with prior work where moderate agreement is considered acceptable in subjective classification tasks (Hallgren, 2012; Sim and Wright, 2005). Given the comparable IRR values between human raters and GPT classifications relative to the human–human baseline, the IRR analysis supports the reliability of the classification

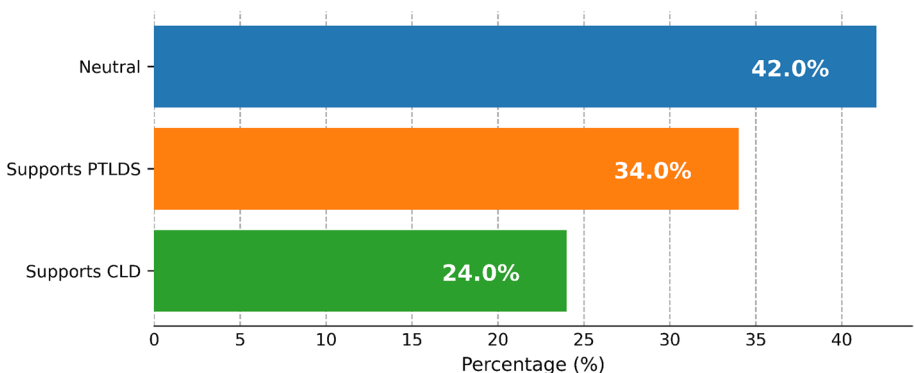


Fig. 6 Percentage distribution of classifications. Neutral, Supports PTLDS, and Supports CLD are depicted in blue, orange and green, respectively

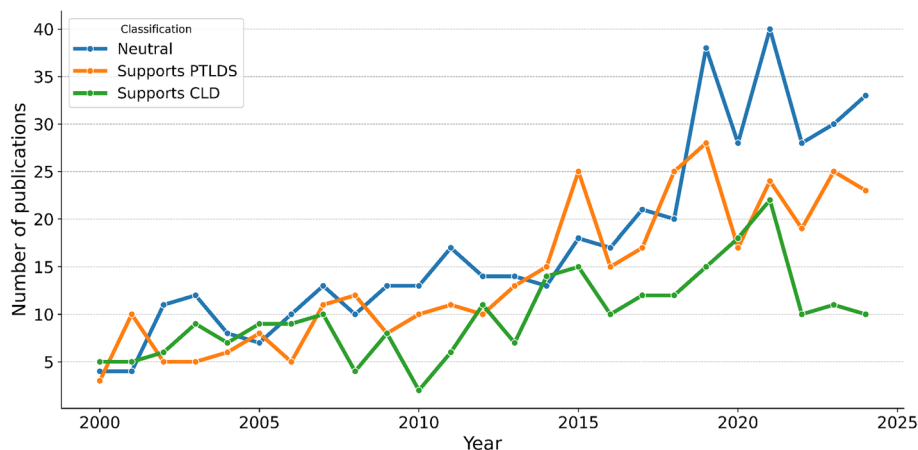


Fig. 7 Study classifications on Lyme disease from 2000 to 2024. The yearly count of abstracts on Lyme disease from 2000 to 2024 was classified into three classifications: Neutral, Supports PTLDS, and Supports CLD, which are depicted as blue, orange, and green, respectively

process and the underlying prompts. Here, reliability should be read as reproducibility of the coding procedure at corpus scale rather than proof that any single classification is definitively correct.

To make the disagreement structure more explicit, we additionally examined the 90 stance-focused abstracts within the validation sample whose classification labels fell into “Supports PTLDS”, “Supports CLD”, or “Neutral”. On this subset, the two human raters agreed on 53 abstracts and disagreed on 37. Most disagreements reflected boundary cases rather than direct opposition: 21 of the 37 involved “Neutral” versus a polar label, 10 involved one rater assigning “Unrelated” while the other assigned a stance label, and only 6 involved a direct “Supports PTLDS” versus “Supports CLD” split. The GPT classification matched at least one human rater in 68 of the 90 cases, including 32 of the 37

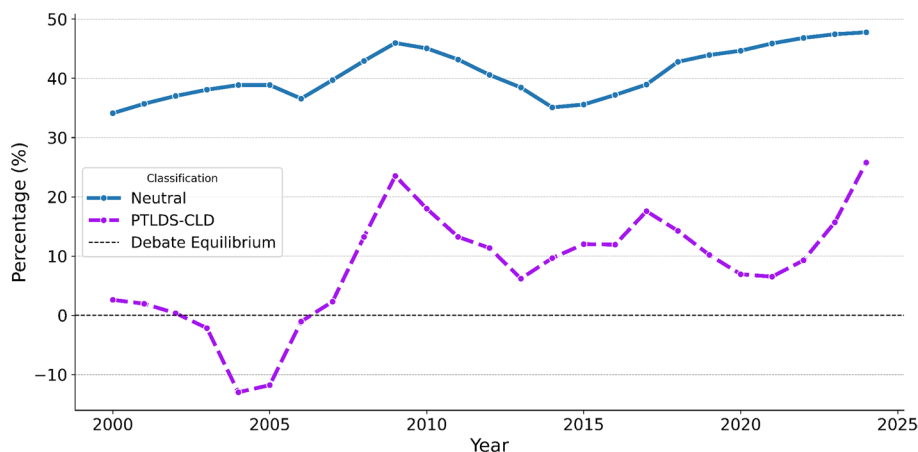


Fig. 8 Smoothed trends in classification percentages by year using the Savitzky–Golay filter. Lines represent smoothed trends for Neutral and the percentage difference of PTLDS minus CLD classifications over the period 2000 to 2024

human-disagreement cases; only 22 cases yielded a third label distinct from both human raters.

Finally, we examined the validity of the justification texts generated by the LLM as rationales for each classification. Human raters achieved a Cohen's Kappa of $\kappa = 0.61$ when selecting between justification options, while both raters achieved $\kappa = 0.71$ against the LLM, indicating that the raters aligned more closely with the LLM-proposed justifications than with each other.

Temporal evolution of competing perspectives (RQ2)

Applying the validated methodology to the 1,033 relevant abstracts, our analysis revealed distinct longitudinal patterns in the Lyme disease discourse. Figure 5 shows an increase in publication volume over time, particularly after 2014. The analysis of stance distribution (Figures 5, 6, 7) reveals that the *Neutral* stance is consistently the largest category (42% overall), consistent with widespread epistemic caution or strategic neutrality in a domain marked by diagnostic and therapeutic uncertainty. However, the combined proportion of polarised studies (*Supports PTLDS* at 34% and *Supports CLD* at 24%) totals 58%, indicating sustained division.

The smoothed trends in Figure 8 show a clear temporal shift (**RQ2**). CLD-supporting studies held a slight relative edge in the early 2000s, but PTLDS-supporting studies became more prominent after 2010. The distribution of published stances therefore shifted over time, with PTLDS-aligned framing becoming more common while CLD perspectives remained a substantial minority.

Venue-level patterns and epistemic gatekeeping (RQ3)

Figure 9 shows the net stance gap across the top 20 journals by publication volume in our dataset, measured as the percentage-point difference between PTLDS- and CLD-supporting studies within each venue. This descriptive measure shows the direction and magnitude of stance skew within individual journals. Figure 10 complements it by showing how far each journal's PTLDS and CLD shares depart from what would be expected given the years in which that journal published. Expected values are calculated from the year-specific

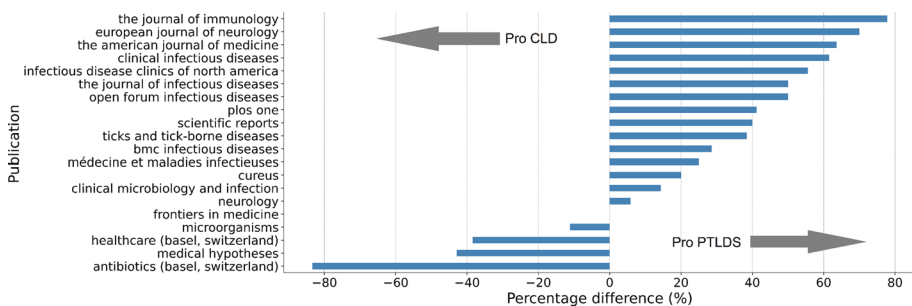


Fig. 9 Top 20 journals by volume representing our dataset and depicting their net stance gap. Values are the difference in percentage points between published papers classified as supporting PTLDS or CLD. Positive values indicate a higher proportion of PTLDS-supporting studies for a journal and negative values indicate a higher proportion of CLD-supporting studies

stance distribution of the full corpus and weighted by each journal's publication-year profile, so the adjusted figure controls for changes in the overall stance mix over time.

Leading infectious disease and clinical medicine journals (e.g., *Clinical Infectious Diseases*, *Journal of Infectious Diseases*, *American Journal of Medicine*) show a higher proportion of PTLDS-supporting studies relative to CLD-supporting studies. CLD perspectives are more frequently observed in a smaller set of niche or hypothesis-driven venues (e.g., *Medical Hypotheses*, *Antibiotics*). Figure 10 shows that this pattern is not simply an artefact of the corpus-wide baseline or of differences in publication-year composition. Several high-status journals remain over-represented for PTLDS and under-represented for CLD relative to these year-adjusted expectations, while a smaller set of niche venues shows the reverse pattern. This venue-level segmentation illustrates how publication patterns shape the visibility of competing knowledge claims and how LLM-derived stance labels can be aggregated into venue-resolved indicators of perspective distribution, consistent with evidence that status and legitimacy in academic publishing shape centre–periphery diffusion across venues (Siler and Larivière, 2024).

Citation dynamics and epistemic hierarchy

We also examined citation impact (RQ3). The top 20 most-cited abstracts, just 2% of the dataset, accounted for 45% of all citations. This concentration indicates that a small subset of publications exerted disproportionate influence. Among these highly cited papers, the

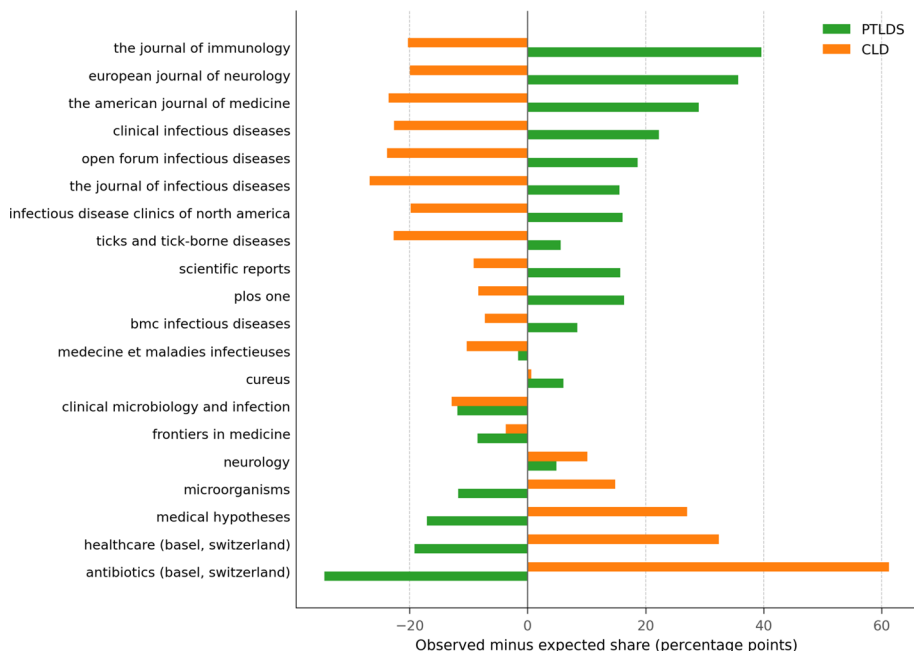


Fig. 10 Time-adjusted over- and under-representation of PTLDS- and CLD-supporting studies across the top 20 journals by related publication volume. Bars show how far the observed percentage-point share for each stance differs from the expected share. Expected values are based on the year-specific stance distribution of the full corpus and weighted by each journal's publication-year profile. Positive values indicate over-representation and negative values indicate under-representation

stance distribution tilted toward PTLDS (13 PTLDS, 3 CLD, 4 Neutral). Across the full dataset, PTLDS-supporting papers made up 34% of studies but received 52% of citations, whereas CLD-supporting papers comprised 24% of studies and received 22% of citations; Neutral papers comprised 42% of studies and received 26% of citations. Citation attention in this corpus therefore falls unevenly across stances, consistent with cumulative advantage in stratified journal systems (Larivière and Gingras, 2010; Siler and Larivière, 2024).

Thematic evolution and information flows (RQ4)

To identify dominant themes within the Lyme disease literature, we categorised 1,033 abstracts based on the core thematic areas derived via Step 4 of the Methodology (see Table 6). This categorisation allows analysis of how research focus and stance are distributed across thematic areas within this contested domain. Table 3 presents this distribution.

The theme *Active Infection vs. Post-Infectious Immune Activity* is the most frequent, appearing in over half the corpus and representing a central axis of disagreement between persistent-infection and post-infectious explanations. This theme is strongly polarised, with most studies supporting one explanatory model or the other. The theme *Diagnostic Complexity and Uncertainty* is also prominent but is predominantly neutral (77%), suggesting substantial attention to diagnostic challenges without adopting an explicit stance. Among studies adopting a stance within this theme, there is a modest preference toward PTLDS-aligned explanations. The theme *Therapeutic Controversies and Antibiotic Efficacy* also shows substantial polarisation, with similar proportions supporting PTLDS and CLD frameworks, reflecting persistent disagreement regarding treatment approaches. Among secondary themes, studies focused on *Immune Dysregulation and Autoimmune Mechanisms* predominantly align with PTLDS (62.5%), while *Mechanisms of Pathogen Persistence and Biofilm Formation* is strongly associated with CLD (90.0%), indicating distinct thematic clusters associated with different explanatory hypotheses.

Neurocognitive and Neuropsychiatric Manifestations is also largely neutral (61.7%), which suggests continuing uncertainty about underlying mechanisms, though stanced papers lean slightly toward PTLDS. *Patient-Centred Experiences and Advocacy* is increasingly visible but remains predominantly neutral (82.6%), which suggests that it has entered the literature without becoming central to the main polarised debate. *Sociocultural and Ethical Factors* is likewise mostly neutral, though the small stanced subset leans toward

Table 3 Distribution of studies by theme and classifications supporting PTLDS/CLD or neutrality.

Theme	Papers	Percent	Neutral	Supports	Supports
			%	PTLDS (%)	CLD (%)
Active infection vs. post-infectious immune activity	579	56.1	11.7	49.4	38.9
Diagnostic complexity and uncertainty	530	51.3	77.0	16.2	6.8
Therapeutic controversies and antibiotic efficacy	365	35.3	20.8	40.5	38.6
Neurocognitive and Neuropsychiatric Manifestations	196	19.0	61.7	21.4	16.8
Immune Dysregulation and Autoimmune Mechanisms	192	18.6	28.1	62.5	9.4
Patient-centred experiences and advocacy	149	14.4	82.6	8.1	9.4
Mechanisms of pathogen persistence and biofilm formation	30	2.9	10.0	0.0	90.0
Sociocultural and ethical factors	25	2.4	76.0	24.0	0.0

PTLDS. Table 4 shows that the literature devotes less attention over time to *Active Infection vs. Post-Infectious Immune Activity* and more to *Diagnostic Complexity and Uncertainty*. Attention to treatment controversies also falls, while work on *Patient-Centred Experiences and Advocacy* grows but remains largely descriptive.

Discussion

With respect to RQ1, the hybrid AI–human framework reproduced expert judgement in complex stance-classification tasks at levels broadly comparable to expert–expert agreement, suggesting sufficient reliability for corpus-level scientometric analysis. The framework combined (i) corpus construction and cleaning, (ii) LLM-based stance and theme classification guided by expert-designed schemas, (iii) expert validation and inter-rater reliability assessment, and (iv) aggregation of classifications into venue-, citation-, and theme-stratified indicators for scientometric analysis, demonstrating its scalability.

Regarding RQ2, our results show a clear longitudinal shift in the published stance distribution. While CLD-aligned studies held a slight relative edge early in the period, PTLDS-aligned studies became increasingly prominent over time, especially after 2010, while Neutral studies remained the single largest category throughout. This temporal pattern indicates not convergence to consensus, but a reweighting of the published discourse within a still-contested field.

For RQ3, our results show that venue structures and citation hierarchies are central to the epistemic configuration of this controversy. Mainstream, high-impact clinical and infectious disease journals predominantly publish PTLDS-aligned articles and account for a disproportionate share of citations, while CLD-aligned articles are concentrated in niche, hypothesis-driven outlets with lower impact and weaker integration into the core citation network. This segmentation means that actors who rely primarily on high-status venues encounter a stance profile in which PTLDS is structurally amplified, whereas CLD perspectives are more likely to remain peripheral and less discoverable within standard scholarly search and indexing channels.

For RQ4, the thematic analysis reveals how competing framings are organised and how they shift over time within this domain. Diagnostic uncertainty persists as a large, largely neutral theme, indicating an enduring area of low consensus in the discourse. Mechanistic debates over active infection versus post-infectious immune processes exhibit clear

Table 4 Shifts in the percentage of studies by thematic focus across decades, with the percentages normalised for the current decade due to its incompleteness

Theme	2000s (%)	2010s (%)	2020s (%)
Active infection vs. post-infectious immune activity	33	28	24
Diagnostic complexity and uncertainty	21	25	29
Therapeutic controversies and antibiotic efficacy	20	20	14
Neurocognitive and neuropsychiatric manifestations	11	8	11
Immune dysregulation and autoimmune mechanisms	10	8	10
Patient-centred experiences and advocacy	3	8	10
Sociocultural and ethical factors	1	1	2
Mechanisms of pathogen persistence and biofilm formation	1	2	1

stance-linked design patterns, with PTLDS-aligned work drawing more heavily on RCTs and observational studies, and CLD-aligned work drawing on case reports, series, and mechanistic studies. Emerging attention to patient experience and advocacy appears predominantly in neutral or descriptive form, suggesting that these concerns have entered the literature but have not yet been fully integrated into dominant evidential framings.

Across these research questions, the hybrid AI–human pipeline proved effective because it formalised a division of labour between expert judgement and scalable automated classification. Experts defined the schema, refined the prompts, and validated the outputs, while the LLM enabled consistent application of that framework across a corpus that would be difficult to code manually at similar scale. This is the practical contribution of the workflow: not the replacement of expertise, but the amplification of it into reproducible scientometric indicators.

These findings also clarify what LLM-empowered scientometrics requires in practice. Prompt engineering alone is not enough; the workflow needs explicit schemas and stability checks such as self-reflection to handle the subtleties of framings and complex scientific texts. LLMs can scale analysis, but they must remain anchored to human validation before they are used for evaluative purposes. The study therefore offers the scientometric community a validated path beyond metadata-only indicators toward semantic mapping of contested scientific literatures at scale.

Limitations

The framework is limited by abstract-only evidence, English-language coverage, residual interpretive ambiguity even after expert validation, and the compression inherent in thematic labels. Abstract-only analysis can miss methodological detail and rhetorical structure in full texts. Database coverage and abstract availability shape the corpus and can bias observed stance prevalence. Some temporal change may reflect shifting terminology rather than changing substantive claims. Because the structure of irretrievable abstracts was not preserved at sufficient granularity for retrospective audit, our Lyme-specific findings should be interpreted as patterns within the retrieved corpus rather than exact prevalence estimates for the full underlying literature. The validation interface also imposed constraints: raters saw two candidate justifications during the task and could not reject both or provide unconstrained free-text alternatives. This design kept the evaluation structured, but it may have introduced anchoring effects and means that the justification results should be interpreted

as comparative judgements within a bounded interface. Lyme disease is also a relatively favourable demonstration case because its controversy is long-running and its stance categories are comparatively legible; other controversies may require broader schema design, multi-label coding, or more iterative prompt development. These limitations are typical of large-scale text-based studies of scholarly communication. Transparent corpus design and empirical validation can mitigate problems of construct validity, representativeness, and indicator robustness, but they cannot remove them entirely (Ding et al., 2013; Thelwall, 2026). Self-reflection is a useful prompting strategy, but prompts remain sensitive to perturbation and can produce variable inference-time results across models. Emerging prompt-optimisation frameworks such as DSPy may improve this stability.

Conclusion

This study introduced and validated a hybrid AI–human method for large-scale analysis of contested scientific discourse. We applied it to a 25-year corpus on the Lyme controversy to map shifts in stance, theme, venue, and citation patterns. The analysis identified temporal changes in published stance distributions and clear asymmetries in venue placement and citation attention. The framework reproduced expert judgement at levels broadly comparable to expert–expert agreement, supporting its use for corpus-level scientometric analysis. In this corpus, work aligned with the mainstream post-treatment syndrome framing (PTLDS), which explains persistent symptoms without positing ongoing active infection after standard therapy, was more prevalent in higher-impact journals and received a disproportionate share of citations. Work aligned with the chronic-infection framing (CLD), which argues for persistent infection and often supports extended antimicrobial treatment, was concentrated in a different segment of the publication landscape. Beyond this specific case, the workflow yielded reproducible stance and theme indicators that could be stratified by venue and citation impact. It offers a scalable and reliable framework for literature-based scientometrics and makes the structure of scientific disagreement measurable rather than anecdotal.

Classification prompts and example outputs

Below is the prompt used for pre-screening classification GPT-4o-mini to execute the Step 2a of the methodology depicted in Figure 2 in the main manuscript.

You are a world-leading expert in medical literature, specialising in Lyme disease and the controversies surrounding chronic Lyme disease (CLD) and post-treatment Lyme disease syndrome (PTLDS).

Your task is to evaluate a single scientific paper abstract at a time to determine if it has relevance to either CLD or PTLDS for further processing.

Definitions:

- Chronic Lyme Disease (CLD): A contested condition wherein some believe symptoms result from an ongoing *Borrelia burgdorferi* infection is often cited as requiring extended antibiotic therapy.
- Post-Treatment Lyme Disease Syndrome (PTLDS): Refers to persistent symptoms (e.g., fatigue, pain, cognitive issues) following standard Lyme disease treatment, generally understood not to involve active infection.

Categories:

Each abstract should be categorised into one of the following three categories:

1. Potentially Related to CLD/PTLDS:
 - Select this category if the paper involves human studies and contains explicit or implicit references, findings or terminology potentially related to CLD or PTLDS. This includes abstracts where the connection to CLD or PTLDS is indirect, uncertain, or requires additional processing to determine relevance.
2. Definitely Unrelated:
 - Select this category if the study has no relevance to CLD or PTLDS in human subjects. Abstracts discussing other medical conditions and unrelated tick-borne diseases.
3. Animal Study:
 - Select this category if the study is entirely focused on animal subjects with no direct relevance to human cases of CLD or PTLDS.

Input:

For each abstract, you will receive an index number, a paper title, and the abstract text.

JSON Output Structure:

Your output must be a JSON object containing only the evaluation of the current abstract. Each response should include the following fields:

- 'index': The index number of the abstract, which MUST match the input.
- 'classification': One of "Potentially Related to CLD/PTLDS," "Definitely Unrelated," or "Animal Study."
- 'confidence': Confidence level in the classification ("High," "Medium," or "Low").

Example Output:

```
{
  "index": 700,
  "classification": "Potentially Related to CLD/PTLDS",
  "confidence": "High"
}
```

ABSTRACT:

< input >

After this classification stage, stance-framing detection was executed as outlined in Step 2b of the methodology in Figure 2 in the main manuscript. The abridged prompt appears below. For brevity, the three abstracts provided to the LLM as part of few-shot learning are omitted here but are available on the project's GitHub repository:

You are a world-leading expert in medical literature analysis and particularly in medical literature, specialising in Lyme disease and the debates surrounding chronic Lyme disease (CLD) and post-treatment Lyme disease syndrome (PTLDS). Your task is to evaluate a series of scientific paper abstracts related to Lyme disease and determine each abstract's stance on PTLDS and CLD. This task requires scientific impartiality to minimize any bias in classifying abstracts concerning disputed subjects such as CLD or PTLDS. The current literature on this topic is polarised and can be expressed as follows:

"Some medical experts argue that post-treatment symptoms, experienced by a subset of patients after completing standard antibiotic therapy, can be attributed to what they define as Post-Treatment Lyme Disease Syndrome (PTLDS). According to some health organisations, these symptoms|ranging from fatigue to cognitive impairment|are likely caused by immune responses or tissue damage, rather than persistent infection. PTLDS concerns patients who experience persistent symptoms for at least six months after the completion of recommended treatment for Lyme disease.

Conversely, other organisations advocate for the recognition of chronic Lyme disease (CLD), contending that ongoing infection or immune dysfunction may be responsible for these symptoms and recommend extended antibiotic regimens, pointing to contested evidence of patient improvement. CLD is a broader term that encompasses a range of unexplained symptoms that patients attribute to a persistent Lyme disease infection that survives recommended treatment. Both PTLDS and CLD involve persistent symptoms, such as fatigue, pain, and neurological issues, and both face with diagnostic difficulties due to nonspecific symptoms and lack of clear biomarkers, contributing to ongoing debates in the medical community."

Based on the above context and the debate, your task is to classify each abstract and determine each abstract's explicit or implicit stance regarding CLD or PTLDS based on the following classifications:

- Supports PTLDS
- Supports CLD
- Neutral
- Unrelated
- Animal Study

Definitions for Classification:

1. Supports PTLDS:
 - Attributes persistent symptoms after successful Lyme disease treatment to other factors like immune dysfunction, chronic fatigue syndrome, or depression.
 - Opposes the use of prolonged or repeated antibiotic treatments for PTLDS.
 - Suggests or implies that CLD lacks scientific support.
2. Supports CLD:
 - A contested condition wherein some believe symptoms result from an ongoing or active *Borrelia burgdorferi** infection after standard treatment, often cited as requiring extended antibiotic therapy.
 - Suggests that persistent symptoms may require prolonged or repeated antibiotic treatment.
3. Neutral:
 - Presents balanced arguments without clearly supporting or refuting either PTLDS or CLD.
 - Discusses aspects of Lyme disease that relate to both sides of the PTLDS and CLD debate without taking a clear stance.
 - No clear stance or support or refutation is provided regarding PTLDS or CLD, either explicitly or implicitly.

4. Unrelated:

- The abstract does not mention PTLDS or CLD, or the surrounding debate at all.
- Focuses on other aspects of Lyme disease, such as acute Lyme disease, vector control, ecological studies, epidemiology, or diagnostic methods without mentioning chronic conditions.
- References CLD without any explicit or implied previous treatment.

5. Animal Study:

- The abstract pertains exclusively to animal models or non-human subjects without direct implications for PTLDS or CLD in humans.
- Does not involve human clinical data or direct conclusions about PTLDS or CLD in humans.
- If the animal study provides evidence directly relevant to PTLDS or CLD in humans, classify it accordingly under "Supports PTLDS", "Supports CLD", or "Neutral".

Additionally, assign a Confidence Level to each classification you make: High, Medium, or Low, based on how confident you are in the given classification.

- High: The chosen classification is indisputably true and cannot possibly be any of the other options.

The abstract contains explicit wording relating to its stance.

- Medium: The chosen classification requires a mix explicit and implicit evidence in the abstract. There is some degree of certainty, but the classification is inferred rather than explicitly stated. Typically arises from indirect language, ambiguous terms, or implied support for a stance.

- Low: The classification is uncertain due to lack of clear evidence or conflicting information in the abstract, relying more on inference and contextual understanding. Typically occurs when the abstract is vague, uses ambiguous language, or includes evidence that could be interpreted in multiple ways. Reflects significant uncertainty in the classification, but that the chosen classification is more likely than the others.

Also, provide a reason comprising 2-3 sentences for your classification above. Clearly state your justification and if the abstract implicitly or explicitly supports PTLDS or CLD in humans. When determining the abstract's stance, consider both explicit statements and implications of the abstracts given context and chosen topic. Look for language that suggests support or opposition even if not directly stated.

Input:

For each abstract, you will receive an index number of the abstract, a paper title, and the abstract text.

JSON Output Structure:

Your output must be a JSON object containing the following fields:

- 'index': The index number of the abstract, which MUST match the input.
- 'classification': One of "Supports PTLDS", "Supports CLD", "Neutral", "Unrelated", or "Animal Study"
- 'confidence': Confidence level in the classification ("High", "Medium", or "Low").
- 'reason': Provide a 2-3 sentence justification for your classification.

< few-shot learning examples comprising 3 abstracts and their classification >

ABSTRACT:

< input >

Inter-rater reliability web application

Web App with example texts and outputs

Figure 11 illustrates our custom-built web application created specifically for the human validation process (Step 3 in Figure 2 in the main manuscript). In this interface, a user (in this case, a subject-expert rater) logs in to evaluate a given abstract. The application displays the original abstract text and prompts the user to choose one of several classification labels (e.g., Supports PTLDS, Supports CLD, Neutral, or Animal Study). Next, the rater selects a confidence level in their choice (e.g., Low, Medium, or High) and selects one of the two options that best describes the justification for the classification.

Figure 12 depicts another view of the same web application but for a different abstract. This example emphasizes the classification process itself, highlighting how a user can read through the abstract, reflect on the core content (e.g., whether it focuses on CLD, PTLDS, or otherwise), and then decide upon the appropriate classification label. The interface

Lyme Classifier
Classifier
Hello user1@lyme.nz!
Log off

Classifier

Article 104 out of 181

[See full List](#)

Abstract

We previously showed that numerous polymorphonuclear leukocyte (PMN) granule components efficiently kill *Borrelia burgdorferi*, the agent of Lyme disease. In addition, motile, granule-poor cytoplasts (U-Cyt) from human blood PMN can exert anti-*Borrelia* activity against opsonized *B. burgdorferi* independently of oxidative mechanisms. Here we show that lysates of U-Cyt also possess anti-*Borrelia* activity, a portion of which comes from the abundant cytosolic protein calprotectin. The anti-*Borrelia* activity of U-Cyt lysates and recombinant calprotectin was partially or completely reversed by specific antibody to calprotectin and by Zn(2+), a cation essential for the growth of *B. burgdorferi* and known to inhibit the antimicrobial activity of calprotectin. Quantitative microscopic and regrowth assays revealed that calprotectin acted in a bacteriostatic fashion against *B. burgdorferi*. We conclude that calprotectin, a potent bacteriostatic agent from a cell primarily recognized for its oxidative and granular antibacterial mechanisms, may play a modulatory role in infection by the Lyme spirochete, particularly at sites of acute inflammation.

Classification

Neutral
Supports PTLDS
Supports CLD
Unrelated
Animal Study

Confidence

Low
Medium
High

Justification

Reasoning1

Upon reevaluation, while the abstract focuses on healthcare providers' knowledge and experiences, this could be interpreted as indirectly relevant to the PTLDS/CLD debate by highlighting the challenges in communicating with patients about chronic Lyme.

Reasoning2

Upon reevaluation, the abstract does not specifically address PTLDS or CLD, nor does it focus on animal studies. Instead, it discusses the bacteriostatic properties of calprotectin against *Borrelia burgdorferi*, which is relevant to the understanding of Lyme disease but does not engage with the chronic aspects of the disease or the debates surrounding PTLDS and CLD. Therefore, it is more accurately classified as 'Unrelated' to the PTLDS and CLD discussions, with a Medium confidence level due to the abstract's focus on a specific protein's role in infection rather than chronic disease outcomes.

Option 1

Option 2

Save

Fig. 11 IRR web app showing the user abstract classification interface

Lyme Classifier

Classifier

Hello user1@lyme.nz! Log off

Classifier

Article 106 out of 181

[See full list](#)

Abstract

Lyme disease (LD) is most often diagnosed clinically, and the differential diagnosis of erythema migrans may be challenging. Recent advances have raised questions about the efficacy of traditional diagnostic modalities, but may soon facilitate consistent identification of patients with Lyme borreliosis. Therapeutic recommendations vary with the stage of disease, and treatment usually leads to complete resolution. The management of patients with "chronic Lyme disease" is controversial. A number of preventative measures have been evaluated; those involving the avoidance of tick bites with protective clothing and insect repellents remain the simplest and most effective. © 2010 by the American Academy of Dermatology, Inc.

Classification

Neutral

Supports PTLDS

Supports CLD

Unrelated

Animal Study

Confidence

Low

Medium

High

Justification

Reasoning1

Upon reevaluation, the abstract maintains a neutral stance as it discusses the management and prevention of Lyme disease without endorsing either PTLDS or CLD. It acknowledges the controversy surrounding 'chronic Lyme disease' but does not provide evidence or language that would suggest support for either side. The mention of treatment leading to complete resolution further reinforces the neutral position, as it does not imply ongoing infection or persistent symptoms. Therefore, the classification remains appropriate, but the confidence level is adjusted to High due to the clarity of the abstract's neutral position.

Reasoning2

Upon reevaluation, the abstract does not maintain a neutral stance; instead, it implicitly leans toward supporting CLD by acknowledging the controversy surrounding 'chronic Lyme disease' while not explicitly dismissing the possibility of ongoing infection or persistent symptoms after treatment. The mention of treatment leading to complete resolution could be interpreted as a dismissal of the experiences of patients who do not achieve complete resolution, subtly undermining the PTLDS perspective. Therefore, the classification is inappropriate and the confidence level should be adjusted to Low due to the potential for misinterpretation of the abstract's stance.

Option 1

Option 2

Save

2025 Lyme Classifier

Fig. 12 IRR web app demonstrating the classification process

displays the choices that a user has selected in green. Together, both figures demonstrate the workflow that expert human raters followed in the IRR analysis process to validate the AI-based stance detection and justification texts provided for the rationale behind the LLM's decisions.

Example classification results post self-reflection

A selection of abstracts whose classification changed from one category to the next post self-reflection prompting together with the justification texts, are shown in Table 5.

Full definitions and descriptions of consolidated themes

Table 6 provides a deeper exposition on the meaning and medical significance of each extracted theme, as well as its social science implications viewed from the Science and Technology Studies framework.

 Springer

Table 5 Selected examples of abstracts with original and revised classifications and justifying text

Abstract	Classifications	Original justification	Revised justification
Background: IL-26 has demonstrated antimicrobial properties, as well as in the degradation of DNA from the Lyme disease spirochete <i>Borrelia burgdorferi</i> (Bb). Additionally, IL-26 can promote macrophage activation and enhance Bb phagocytotic activity. It is unclear if cell-mediated immune responses are modulated through TLR9 signaling when exposed to IL-26 Bb DNA complexes in post-treatment Lyme disease syndrome (PTLDS). Objective: We here aim to explore the effect of IL-26 in human Toll-like receptor (TLR)-9's activation upon the recognition of Bb DNA. Methods: We utilized a single-receptor cell system, HEK-DualTM hTLR9 cells, which harbors two reporter plasmids for the NF-κB and IL-8 signaling pathways. Bb DNA was exposed to increasing concentrations of IL-26 in monomeric or dimeric form and then used to stimulate the cells for 4 h. The TLR-9 ligand CpG was used as a control. Results: We observed that NF-κB and IL-8 activation was maximal when the cells were stimulated with Bb DNA that had been treated with 5 μM of IL-26 monomer and 1 μM of IL-26 dimer. The same was observed for IL-8 activation upon CpG stimulation. We observed, however, a decrease in NF-κB activation when treated with either form of IL-26. An NF-κB activation increase did not occur with IL-26-treated TLR9 ligand CpG. Conclusions: Our study shows an enhancement in NF-κB and IL-8 activation upon the recognition of IL-26-treated Bb DNA by TLR9, which suggests an increase in sensing by the TLR9 of Bb DNA when it is in the form of an IL-26-Bb DNA complex. These findings will prompt further studies on the interaction between IL-26 and Bb DNA. 145	Original: Neutral - Revised: Supports PTLDS	The abstract discusses the role of IL-26 in the immune response to <i>Borrelia burgdorferi</i> DNA and its potential implications for post-treatment Lyme disease syndrome (PTLDS). However, it does not take a clear stance on whether persistent symptoms are due to ongoing infection or other factors, nor does it advocate for or against the concepts of PTLDS or chronic Lyme disease (CLD). The focus is primarily on the biochemical interactions and immune mechanisms without explicit support for either side of the debate.	Upon reevaluation, the abstract explicitly mentions 'post-treatment Lyme disease syndrome (PTLDS)' and explores immune responses related to this condition. The focus on IL-26's role in enhancing immune sensing of Bb DNA suggests a mechanism that aligns with the PTLDS perspective, which attributes persistent symptoms to immune dysfunction rather than ongoing infection. This indicates a clear inclination towards supporting PTLDS, justifying a revision to 'Supports PTLDS' with a High confidence level.

Table 5 (continued)

Abstract	Classifications	Original justification	Revised justification
Arthritis is the most common late manifestation of <i>Borrelia burgdorferi</i> infection in the United States, usually beginning months after the tick bite. In most patients with Lyme arthritis (LA) today, arthritis is the presenting manifestation of the disease. Patients have swelling and pain in one or a few large joints, especially the knee. Serologic testing is the mainstay of diagnosis. Responses to antibiotic treatment are generally excellent, although a small percentage of patients have persistent, postinfectious synovitis after 2 to 3 months of oral and IV antibiotics, which respond to anti-inflammatory therapies. Herein we review the clinical presentation, diagnosis, and management of LA. 7787	Original: Unrelated - Revised: Neutral	The abstract focuses on Lyme arthritis as a manifestation of <i>Borrelia burgdorferi</i> infection, discussing its clinical presentation, diagnosis, and management. While it mentions persistent postinfectious synovitis, it does not engage with the debates surrounding PTLDS or CLD, nor does it imply a stance on these conditions. The content is primarily concerned with the acute and post-treatment aspects of Lyme arthritis rather than chronic conditions.	Upon reevaluation, the abstract does mention persistent postinfectious synovitis, which could be interpreted as a connection to ongoing symptoms after treatment, potentially linking it to the PTLDS debate. However, it does not explicitly support or refute either PTLDS or CLD, making it more appropriate to classify it as Neutral rather than Unrelated. The confidence level is adjusted to Medium, as the abstract does touch on relevant themes but lacks a clear stance.
The immuno-pathogenetic mechanisms underlying chronic Lyme neuroborreliosis are mainly unknown. Human <i>Borrelia burgdorferi</i> (Bb) infection is associated with Bb-specific secretion of interferon- γ (IFN- γ), which may be important for the elimination of Bb, but this may also cause tissue injury. In order to increase the understanding of the pathogenic mechanisms in chronic neuroborreliosis, we investigated which cell types that secrete IFN- γ . Blood mononuclear cells from 13 patients with neuroborreliosis and/or acrodermatitis chronica atrophicans were stimulated with Bb antigen and the phenotypes of the induced IFN- γ -secreting cells were analyzed with three different approaches. Cells expressing CD8 or TCR $\gamma\delta$, which both have cytolytic properties, were the main phenotypes of IFN- γ -secreting cells, indicating that tissue injury in chronic neuroborreliosis may be mediated by cytotoxic cells.	Original: Supports CLD - Revised: Neutral	The abstract discusses the immunological response in patients with chronic Lyme neuroborreliosis, suggesting that the secretion of interferon- γ by specific cell types may contribute to tissue injury. This aligns with the CLD perspective, which posits that ongoing immune responses and potential persistent infection could lead to chronic symptoms. However, the abstract does not explicitly advocate for prolonged antibiotic treatment or directly state that ongoing infection is the cause of symptoms, leading to a medium confidence classification.	Upon reevaluation, the abstract primarily focuses on the immunological mechanisms involved in chronic Lyme neuroborreliosis without explicitly supporting the notion of ongoing infection or the need for prolonged antibiotic treatment. While it discusses the role of immune responses in tissue injury, it does not provide clear evidence that aligns with the CLD perspective. Therefore, it is more accurately classified as Neutral, as it does not take a definitive stance on either PTLDS or CLD.

Table 6 Key overarching themes extracted from the dataset of abstracts, their classifications, and their definition and grounding in the science and technology studies framework

Final consolidated theme	Description	Social science rationale/Relevance
Active infection vs. post-infectious immune activity	Whether persistent symptoms arise from ongoing <i>Borrelia</i> infection or post-infectious immune response.	<i>Paradigm Conflict</i> : Captures the central epistemic divide between infectious and immunological models. Illustrates STS perspectives on controversy and scientific legitimacy.
Diagnostic complexity and uncertainty	Diagnostic ambiguity due to serological limitations and absent biomarkers.	<i>Constructed Diagnosis & Medical Ambiguity</i> : Reflects the social shaping of diagnosis and patient struggles with uncertainty. Grounded in medical sociology and STS views on diagnostic exclusion.
Therapeutic controversies and antibiotic efficacy	Disputes over prolonged antibiotics, evidence standards, and competing guidelines.	<i>Bioethics & Medical Pluralism</i> : Raises ethical tensions between clinical authority, patient autonomy, and advocacy. Relevant to bioethics and health decision-making.
Immune dysregulation and autoimmune mechanisms	Symptoms attributed to immune dysfunction, autoimmunity, or persistent inflammation.	<i>Biomedical Framing & Paradigm Shift</i> : Signals a move from infection models to immunological explanations. Aligns with STS studies of biomedical paradigm evolution.
Neurocognitive and neuropsychiatric manifestations	Cognitive, neurological, and psychiatric symptoms associated with Lyme disease.	<i>Psychosocial Burden & Stigma</i> : Highlights contested mental health framings and stigma in chronic illness. Relevant to sociology of health and chronicity.
Patient-centred experiences and advocacy	Patient narratives, diagnostic dismissal, and lay knowledge claims.	<i>Patient Agency & Resistance</i> : Reflects challenges to institutional medicine and rise of experiential expertise. Links to advocacy and digital health scholarship.
Sociocultural and ethical factors	Role of media, advocacy, and legal framing in public discourse.	<i>Policy & Sociocultural Framing</i> : Shows how media and law shape perceptions and policy agendas. Anchored in STS, media studies, and political sociology.
Mechanisms of pathogen persistence and biofilm formation	Hypotheses of microbial survival mechanisms underlying treatment failure.	<i>Marginalised Science & Alternative Paradigms</i> : Represents CLD-aligned biomedical research challenging dominant views. Connects to STS work on epistemic exclusion and heterodox science.

Author contributions Conceptualization: LG, TS; Methodology: TS, CP; Data curation: CP, TZ, NA; Formal analysis and investigation: TS, CP; Validation: KG, LG, CP; Visualization: NA, TS; All authors contributed to Writing - original draft preparation and Writing - review and editing; Project administration: TS. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. None to declare.

Data availability All data not subject to copyright restrictions, prompts, and outputs are openly available on the GitHub repository <https://github.com/teosusnjak/Lyme-disease-controversy>.

Declarations

Conflict of interest/Competing interests The authors have no relevant financial or non-financial interests or other competing interests to disclose.

Declaration of generative AI in scientific writing The use of generative AI in manuscript preparation is disclosed as follows. Several frontier LLMs from Google, OpenAI, and X served as dialogic partners and editorial assistants during drafting and revision. Their roles included (i) critiquing drafts, (ii) helping refine arguments and the theoretical framing, (iii) suggesting relevant literature, and (iv) improving clarity, tone, and structure. The authors retained full intellectual control throughout and remain fully responsible for the final content of the publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alturayef, N., Luqman, H., & Ahmed, M. (2023). A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7), 5113–5144. <https://doi.org/10.1007/s00521-023-08285-7>
- Arhiliuc, C., Guns, R., Daelemans, W., & Engels, T. C. (2025). Journal article classification using abstracts: a comparison of classical and transformer-based machine learning methods. *Scientometrics*, 130(1), 313–342. <https://doi.org/10.1007/s11192-024-05217-7>
- Bharti, P. K., Dalal, V., & Panchal, M. (2026). Co-reviewer: can ai review like a human? an agentic framework for llm-human alignment in peer review. *Scientometrics*, 1–42. <https://doi.org/10.1007/s11192-026-05557-6> Retrieved from <https://doi.org/10.1007/s11192-026-05557-6>
- Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ Open*, 7(2), Article e012545. <https://doi.org/10.1136/bmjopen-2016-012545>
- Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1), 224. <https://doi.org/10.1057/s41599-021-00903-w>
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662–679. <https://doi.org/10.1080/1369118X.2012.678878>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

- Cabanac, G., & Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, 72(12), 1461–1476. <https://doi.org/10.1002/asi.24495>
- Cameron, D. J., Johnson, L. B., & Maloney, E. L. (2014). Evidence assessments and guideline recommendations in lyme disease: the clinical management of known tick bites, erythema migrans rashes and persistent disease. *Expert review of anti-infective therapy*, 12(9), 1103–1135. <https://doi.org/10.1586/14787210.2014.940900>
- Cetina, K. K. (1999). *Epistemic cultures: How the sciences make knowledge*. harvard University Press.
- Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2), 206–219. <https://doi.org/10.1197/jamia.m1929>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Costas, R., Zahedi, Z., & Wouters, P. (2015). Do altmetrics correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the association for information science and technology*, 66(10), 2003–2019. <https://doi.org/10.1002/asi.23309>
- Dadkhah, M., Oermann, M. H., Hegedüs, M., Raman, R., & Dávid, L. D. (2023). Detection of fake papers in the era of artificial intelligence. *Diagnosis*, 10(4), 390–397. <https://doi.org/10.1515/dx-2023-0090>
- Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, 7(3), 583–592. <https://doi.org/10.1016/j.joi.2013.03.003>
- Dumes, A. A. (2020). *Divided bodies: Lyme disease, contested illness, and evidencebased medicine*. Duke University Press.
- Elliott, K. C. (2021). The value-ladenness of transparency in science: lessons from lyme disease. *Studies in History and Philosophy of Science*, 88, 1–9. <https://doi.org/10.1016/j.shpsa.2021.03.008>
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4), 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>
- Fairclough, N. (1992). *Discourse and social change*. Polity Press.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Gera, P., & Neal, T. (2025). Deep learning in stance detection: A survey. *ACM Computing Surveys*, 58(1), 1–37. <https://doi.org/10.1145/3744641>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), Article e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure [arXiv:2203.05794](https://arxiv.org/abs/2203.05794) arXiv preprint.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. <https://doi.org/10.20982/tqmp.08.1.p023> Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/>
- Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, 65(4), 656–669. <https://doi.org/10.1002/asi.23101>
- Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems*, 34, 2018–2033.
- Khiabani, P. J., & Zubiaga, A. (2025). Cross-target stance detection: A survey of techniques, datasets, and challenges. *Expert Systems with Applications*, 283, Article 127790. <https://doi.org/10.1016/j.eswa.2025.127790>
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
- Küçük, D., & Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1), 1–37. <https://doi.org/10.1145/3369026>
- Lan, X., Gao, C., Jin, D., & Li, Y. (2024). Stance detection with collaborative roleinfused llm-based agents. *Proceedings of the international aaai conference on web and social media (Vol. 18, pp. 891–903)*. <https://doi.org/10.1609/icwsm.v18i1.31360>
- Lantos, P. M. (2011). Chronic lyme disease: the controversies and the science. *Expert review of anti-infective therapy*, 9(7), 787–797. <https://doi.org/10.1586/eri.11.63>
- Larivière, V., & Gingras, Y. (2010). The impact factor's matthew effect: A natural experiment in bibliometrics. *Journal of the American society for information science and technology*, 61(2), 424–427. <https://doi.org/10.1002/asi.21232>

- Latour, B., & Woolgar, S. (2013). *Laboratory life: The construction of scientific facts*. Princeton University Press.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., & Kang, J. (2020). Biobert: a pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36, 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Li, L., Mathrani, A., & Susnjak, T. (2025a). Automated risk-of-bias assessment of randomized controlled trials: A first look at a gepa-trained programmatic prompting framework. arXiv preprint [arXiv:2512.01452](https://arxiv.org/abs/2512.01452)
- Li, L., Mathrani, A., & Susnjak, T. (2025). Transforming evidence synthesis: A systematic review of the evolution of automated meta-analysis in the age of ai. *Research Synthesis Methods*, 1–48. <https://doi.org/10.1017/rsm.2025.10065>
- Li, L., Mathrani, A., & Susnjak, T. (2025). What level of automation is good enough? a benchmark of large language models for meta-analysis data extraction. *Research Synthesis Methods*, 1–22. <https://doi.org/10.1017/rsm.2025.10066>
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282. <https://doi.org/10.11613/bm.2012.031> Retrieved from <https://pubmed.ncbi.nlm.nih.gov/23092060/>
- Mu, R., & Wang, H. (2022). A systematic literature review of open innovation in the public sector: comparing barriers and governance strategies of digital and non-digital open innovation. *Public Management Review*, 24(4), 489–511. <https://doi.org/10.1080/14719037.2020.1838787>
- Murray, D., Ni, C., Gu, W., & Hubbard, T. (2025). Using language models to label clusters of scientific documents. *Scientometrics*, 1–30. <https://doi.org/10.1007/s11192-025-05445-5>
- Olechnowicz, C., Leahy, J., Gardner, A., & Sponarski, C. C. (2023). Perceived vulnerability for lyme disease questionnaire: A social science tool for understanding tick-borne disease attitudes. *Ticks and Tick-borne Diseases*, 14(2), Article 102120. <https://doi.org/10.1016/j.ttbdis.2023.102120>
- Ortega, J. L. (2022). Classification and analysis of pubpeer comments: How a web journal club is used. *Journal of the Association for Information Science and Technology*, 73(5), 655–670. <https://doi.org/10.1002/asi.24568>
- Peretti-Watel, P., Ward, J., Lutaud, R., & Seror, V. (2019). Lyme disease: Insight from social sciences. *Médecine et Maladies Infectieuses*, 49(2), 133–139. <https://doi.org/10.1016/j.medmal.2018.12.005>
- Puppo, C., Hansmann, Y., Moïnot, L., Duval, X., Chirouze, C., & Préau, M. (2023). The social representations of diagnosing lyme disease. *Plos one*, 18(2), Article e0276800. <https://doi.org/10.1371/journal.pone.0276800>
- Rebman, A. W., Bechtold, K. T., Yang, T., Mihm, E. A., Soloski, M. J., Novak, C. B., & Aucutt, J. N. (2017). The clinical, symptom, and quality-of-life characterization of a well-defined group of patients with posttreatment lyme disease syndrome. *Frontiers in Medicine*, 4, 224. <https://doi.org/10.3389/fmed.2017.00224> (PMID: 29312942; PMID: PMC5735370)
- Renze, M., & Guven, E. (2024). Self-reflection in llm agents: Effects on problemsolving performance. 2024 2nd International Conference on Foundation and Large Language Models (FLLM), 516–525. <https://doi.org/10.1109/fllm63129.2024.10852426>
- Sawaki, T. (2014). On the function of stance-neutral formulations: Apparent neutrality as a powerful stance constructing resource. *Journal of English for Academic Purposes*, 16, 81–92. <https://doi.org/10.1016/j.jeap.2014.10.001>
- Schreier, M. (2012). Qualitative content analysis in practice.
- Siler, K., & Larivière, V. (2024). Varieties of diffusion in academic publishing: How status and legitimacy influence growth trajectories of new innovations. *Journal of the Association for Information Science and Technology*, 75(2), 132–151. <https://doi.org/10.1002/asi.24844>
- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257–268. <https://doi.org/10.1093/ptj/85.3.257> Retrieved from <https://academic.oup.com/ptj/article/85/3/257/2805022>
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3), 943–950. <https://doi.org/10.1038/s41591-024-03423-7>
- Stricker, R. B., & Johnson, L. (2008). Chronic lyme disease and the ‘axis of evil’. *Future Microbiology*, 3(6), 621–624. <https://doi.org/10.2217/17460913.3.6.621>
- Susnjak, T. (2024). Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. *Borrelia burgdorferi: Methods and protocols* (pp. 173–183). Springer. https://doi.org/10.1007/978-1-0716-3561-2_14
- Thelwall, M. (2026). Large language models and responsible research evaluation: an extension of the leiden manifesto. *Scientometrics*, 1–23. <https://doi.org/10.1007/s11192-026-05552-x>
- Tonks, A. (2007). Lyme wars. *BMJ*, 335(7626), 910–912. <https://doi.org/10.1136/bmj.39363.530961.ad>

- Ullah, M., Shahid, A., Din, I. U., Roman, M., Assam, M., Fayaz, M., & Aljuaid, H. (2022). Analyzing interdisciplinary research using co-authorship networks. *Complexity*, 2022(1), 2524491. <https://doi.org/10.1155/2022/2524491>
- Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7), 5731–5780. <https://doi.org/10.1007/s10462-022-10144-1>
- Wormser, G. P., Dattwyler, R. J., Shapiro, E. D., Halperin, J. J., Steere, A. C., Klempner, M. S., et al. (2006). The clinical assessment, treatment, and prevention of lyme disease, human granulocytic anaplasmosis, and babesiosis: clinical practice guidelines by the infectious diseases society of america. *Clinical Infectious Diseases*, 43(9), 1089–1134. <https://doi.org/10.1086/508667>
- Xie, Q., & Waltman, L. (2025). A comparison of citation-based clustering and topic modeling for science mapping. *Scientometrics*, 130(5), 2497–2522. <https://doi.org/10.1007/s11192-025-05324-z>
- Yousif, A., Niu, Z., Tarus, J. K., & Ahmad, A. (2019). A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review*, 52(3), 1805–1838. <https://doi.org/10.1007/s10462-017-9597-8>
- Zhang, B., & Chen, Y. (2025). Semantic filtering meets knowledge elements: a novel approach for emerging research topics discovery. *Scientometrics*, 131(1), 291–328. <https://doi.org/10.1007/s11192-025-05476-y>
- Zhang, Y., Jin, R., & Zhou, Y. (2020). Supporting topic modeling and trends analysis in biomedical literature. *Journal of Biomedical Informatics*, 110, Article 103574. <https://doi.org/10.1016/j.jbi.2020.103574>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.