Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainability and Interpretability Methods for Large Language Models: An Exploratory Systematic Mapping Study

Georgios Makridis
gmakridis@unipi.gr
University of Piraeus
George Fatouros
University of Piraeus
Stefanos Kampouris
University of Piraeus
John Soldatos
ExpertAI SARL
Dimosthenis Kyriazis
University of Piraeus

Research Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11394163/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainability and Interpretability Methods for Large Language Models: An Exploratory Systematic Mapping Study

Georgios Makridis[1,2*], George Fatouros[1], Stefanos Kampouris[1], John Soldatos[1], Dimosthenis Kyriazis[1]

[1*]Department of Digital Systems, University of Piraeus, Piraeus, Greece.
[2]ExpertAI SARL, Luxembourg, Luxembourg.

*Corresponding author(s). E-mail(s): gmakridis@unipi.gr;

Abstract

Explainability for large language models (LLMs) spans circuit analysis, feature discovery, attribution, probing and editing, rationale faithfulness, source attribution, benchmarks and human-facing explanation. This systematic mapping study charts studies that explain an LLM or an LLM-centred pipeline, excluding work that utilises an LLM to explain another system. Seven sources were searched for reports from 2019 to August 2026 that explain at least one decoder-only or encoder–decoder model of at least 1B parameters. Of 7,825 records (5,730 after deduplication), 1,252 studies were included and coded by method family, explanandum, required access, evaluation construct and a four-field evidence profile. Screening, eligibility and coding were performed by a disclosed AI reviewer whose every decision carries a verbatim quote, and were validated against a human reviewer: title/abstract agreement reached $\boldsymbol{\kappa = 0.84}$ after two failed calibration rounds, full-text agreement $\boldsymbol{\kappa = 0.72}$, and conservative screening recall was 0.78. Methods requiring weights or activations account for 69.8% of the studies and methods compatible with hosted text or logit access for 23.3%; faithfulness or causal-validity evidence is coded in 70.4% (an approximate AI-coded estimate), human utility in 3.4%, deployed settings in 7 studies and specific regulatory obligations in 10. The map indicates that most explanation methods require access that hosted LLMs do not expose, and that human-centred and deployment evidence remains scarce. The protocol, its amendment and every decision are released.

Keywords: Explainable artificial intelligence, Large language models, Mechanistic interpretability, Faithfulness, Systematic mapping study, AI-assisted systematic review

1 Introduction

The diffusion of LLMs has changed the practical meaning of explainability. Earlier XAI work addressed classifiers with stable inputs, bounded label spaces and well-defined prediction events, while LLMs operate through open-ended generation, prompt-conditioned behaviour, multi-turn interaction, retrieval augmentation, tool use and, increasingly, agentic execution. Their outputs are not only predictions but rationales, plans, code, citations and tool calls; as a result, explainability for LLMs comprises both internal questions about mechanism and external questions about whether generated explanations are faithful, useful and auditable (Jacovi and Goldberg 2020).

In this context, this study is organised around a strict direction boundary. The target literature explains LLM behaviour (i.e. how a model stores, transforms, retrieves or communicates information, which circuits, features, activations, prompts, retrieved passages or reasoning traces matter, and whether an explanation corresponds to the computation that produced the output). A separate literature utilizes LLMs as explanation generators for other systems (e.g. narrating the SHAP values of a tabular classifier); that literature treats the LLM as explanator rather than explanandum, answers a different question, and is excluded.

1.1 Positioning against existing reviews

The growing interest in LLM explainability has prompted several secondary studies, six of which are sufficiently close in scope to require explicit positioning. A related survey comes from Luo and Specia (2024), where local and global analysis, model editing, controlled generation and evaluation methods are organised from understanding to utilization, without a reproducible search window or a per-study matrix. In the same direction, Zhao et al. (2024) review how explainability uncovers the functioning of LLMs, covering mechanistic interpretability, probing, representation engineering, training dynamics and alignment. Cambria et al. (2024) put forward a systematic mapping of the relation between XAI and LLMs, reporting 1,030 initial manuscripts, 233 relevant papers and 35 selected articles, and separating application from discussion papers. Another relevant proposition comes from Herrera (2025), where a human-centred reflection is organised around faithfulness, truthfulness, plausibility and contrastivity, with an emphasis on governance-ready design rather than per-study coding. Palikhe et al. (2025) survey explainable LLMs through an architecture-aware taxonomy (i.e. encoder-only, decoder-only and encoder–decoder models) and tabulate studies by method, model, access, scope and integration. Finally, Gao et al. (2026) present a survey of *intrinsic* interpretability, described by its authors as systematic although no search or selection process is reported, organising intrinsically transparent architectures into five design paradigms (i.e. functional transparency, concept alignment, representational decomposability, explicit modularization and latent sparsity induction); the present study is complementary, since intrinsically interpretable architectures fall under IC2 only when their explanations are evaluated on an LLM of $\geq$1B parameters, and the bulk of the mapped literature concerns post-hoc explanation of trained models.

Table 1 Positioning against the closest adjacent reviews. "Matrix published" indicates whether a per-study classification is provided in a form a reader can audit. Each cell was checked against the cited paper; the page or table that supports it is given in Appendix A.

	Luo and Specia (2024)	Zhao et al. (2024)	Cambria et al. (2024)	Herrera (2025)	Palikhe et al. (2025)	Gao et al. (2026)	This study
Scope	XAI for LLMs, understanding to utilization	How LLMs work, explainability view	XAI–LLM relation	Human-centred XAI for LLMs	Explainable LLMs by architecture	Intrinsic interpretability designs	Post-hoc and intrinsic explanation of LLMs ≥1B
Method	Narrative survey	Narrative overview	Systematic mapping	Reflection and survey	Narrative survey	Survey (self-described systematic; no search reported)	Systematic mapping; PRISMA 2020, PRISMA-S
Search window	Not stated	Not stated	2010 to Oct. 2023 (arXiv); stage counts	Not stated	Not stated	"Up to late 2025"; no counts	2019-01-01 to 2026-08-28
Matrix published	No	No	PDF table (application papers only)	No	PDF table	Paper list	Online Resource 1
Access regime	Not discussed	Discussed	Model-agnostic column; not counted	Implicit only	Column in table	Not applicable	Coded (D3)
Evidence coding	No	No	No	No	No	No	Four-field profile

The main difference of the present work compared to the aforementioned ones is that it reports a reproducible, per-record selection trail and codes every included study along access regime and a non-ordinal evidence profile in a machine-readable matrix (Table 1). Two of the reviews tabulate studies in the paper itself, Palikhe et al. (2025) with an access column and Cambria et al. (2024) for its application papers, but neither codes the evidence that the studies provide. It should be noted that this is a difference in method and coverage of reporting, not a claim of broader topical scope.

1.2 Contributions

In this direction, the contributions of the study can be summarised as follows:

1. A classification framework for XAI-for-LLM studies that covers method family, explanandum granularity, access regime and evaluation construct, together with a four-field evidence profile that separates interventional validity, human-subject evidence, ecological validity and statistical reporting instead of ranking them on one ordinal scale.
2. A census-scale selection and coding procedure with a disclosed AI reviewer, in which every decision carries a verifiable quote and the reviewer is validated against a human reviewer through pre-committed calibration rounds and a recall estimate on its exclusions; failed rounds and the superseded automated pilot are reported, and every script, instruction file and per-record decision is released.
3. A map of 1,252 included studies that quantifies the mismatch between the access regimes that explanation methods require and the access that hosted LLMs expose, and that separates faithfulness, plausibility and human-utility claims.

The remainder of the paper is organized as follows: Section 2 presents the background and the explanandum boundary, while Section 3 states the research questions. Section 4 delivers the methodology, including the disclosed AI reviewer and its validation, and Section 5 the selection, the validation of the AI reviewer and the characteristics of the included studies. Section 6 and Section 7 analyse method families and cross-cutting dimensions, Section 8 discusses gaps and research directions, and Section 9 and Section 10 conclude with limitations and recommendations.

2 Background and Preliminaries

2.1 Why LLM explainability is distinct

LLMs differ from earlier systems in five ways relevant to explanation. The same model performs many tasks, so a local explanation of one answer may not generalise to a stable decision rule; outputs are generated token by token, so an explanation may target a token, a span, a reasoning chain, a retrieved source or a whole interaction; and the input space includes system prompts, user messages, retrieved documents, demonstrations, tools and context. Moreover, many deployed models are API-only, so the white-box methods are unavailable exactly where explanation is most requested, and LLMs can produce explanations in natural language, which creates the temptation to accept a fluent rationale as evidence of why an answer was generated. Considering the latter, a rationale, citation, attention map, circuit, feature dictionary and probe result may all be called explanations, but they should not be judged by the same construct.

2.2 Terminology

XAI is used here as the umbrella for methods that make LLM behaviour understandable, inspectable, attributable or accountable. Interpretability denotes methods that provide a structured account of internal or behavioural mechanism, and mechanistic interpretability is narrower still, seeking explanations in components, circuits, features and causal interventions. Rather than impose a philosophical definition, extraction records how each study operationalises the term.

Table 2 Boundary rules for direction and topic (abridged; see Online Resource 2, Section 1).

Case	Decision
LLM explains its own answer (CoT, self-explanation) and the explanation is evaluated	Include (RF)
LLM narrates SHAP or LIME output of another model	Exclude (X-DIR)
Sparse autoencoders, activation patching, probing of LLM internals ($\geq$1B)	Include
The same methods applied only to models below 1B parameters	Exclude (X-MODEL)
LLM-as-judge scoring explanations of LLMs	Include (BM)
LLM-as-judge scoring explanations of other models, or text quality only	Exclude (X-DIR)
Knowledge editing or activation steering with no localisation or explanatory claim	Exclude (X-TOPIC)
Agent or pipeline system whose trace or source attribution is not evaluated as an explanation	Exclude (X-TOPIC)
Retrieval source attribution and citation grounding	Include (SA)
Vision–language model interpretability	Include if language component $\geq$1B
Encoder-only models (e.g. BERT, RoBERTa) only	Exclude (X-MODEL)
Summarisation or translation "faithfulness" (factual consistency)	Exclude (X-TOPIC)

2.3 The explanandum boundary

Included studies explain an LLM or LLM-centred pipeline, while excluded studies utilize an LLM primarily to explain something else, even where they use the words explainable, interpretable or rationale. Table 2 states the pre-decided boundary rules; the full set, with positive and negative example records, is given in the codebook (Online Resource 2).

3 Research Questions

- **RQ1:** Which method families have been proposed for explaining LLM behaviour?
- **RQ2:** What is explained (i.e. tokens, spans, examples, features, circuits, layers, trajectories or pipelines)?
- **RQ3:** Which access regime does each method require, and how does it relate to the availability of the models evaluated?
- **RQ4:** How is explanation quality evaluated, and how often are faithfulness, plausibility and human utility evaluated separately?
- **RQ5:** What kinds of evidence (interventional, human-subject, ecological, statistical) support the explanatory claims?

- **RQ6:** How often do studies connect explanation outputs to specific transparency or regulatory obligations?

4 Methodology

4.1 Review type, reporting and registration

The study is an exploratory systematic mapping study in the sense of Petersen et al. (2015) and Kitchenham and Charters (2007): it charts what has been studied and how, rather than estimating intervention effects. It is reported in accordance with PRISMA 2020 (Page et al. 2021), with search reporting following PRISMA-S (Rethlefsen et al. 2021); items that do not apply to a mapping study (i.e. effect measures and certainty of evidence) are declared as such in Appendix B. This review was not registered. The original protocol was written as a record of the procedure followed rather than as a pre-commitment, and readers should weight it accordingly. The revised protocol that governs the human verification (`protocol_p8.md`) was frozen on 2026-09-13, before any human screening, and is released with the study, but it was not registered. It was amended on 2026-10-03 (Amendment A1, `protocol_amendment_A1_ai_reviewer.md`), before any census decision was made, to replace the human first reviewer by a disclosed AI reviewer whose decisions are validated against a human reviewer (Section 4.6); every step of the calibration and validation, including the rounds that failed and the rules that were fixed before each round, is logged with its date in the amendment. All deviations from the original protocol are listed in Appendix G.

4.2 Information sources

Seven sources were searched. Scopus provides broad indexing and the deduplication backbone, and arXiv is necessary because much interpretability work appears first as a preprint. The ACL Anthology covers NLP venues that Scopus indexes only partially and that IEEE, Springer and Elsevier do not index at all, IEEE Xplore and the ACM Digital Library cover engineering, HCI and FAccT venues, and ScienceDirect and SpringerLink add publisher coverage at lower precision. Citation searching was planned in the original protocol but has not been executed and is reported as not performed.

4.3 Search strategy

A two-tier union was applied to every source. Tier A combined a population block (e.g. "large language model*", "LLM", "ChatGPT", "LLaMA") with direction-locked vocabulary (e.g. mechanistic interpretability, activation patching, sparse autoencoder, logit lens, steering vector, probing classifier, token attribution) in title, abstract and keyword fields, since these terms presuppose that the model is the object of explanation. Tier B combined the same population block with generic vocabulary (e.g. explainab*, interpretab*, XAI, Shapley, counterfactual explanation) restricted to title and author-keyword fields, because these terms retrieve both directions. The executed

strings, filters, dates, raw and exported yields for every source, including all superseded strategies, are given in Appendix D; the term lists of the locally executed rules are imported directly from the released scripts.

It should be noted that three interfaces required local correction. The arXiv API does not enforce quoted phrase matching in the abstract field, so the intended Boolean was enforced locally (13,685 retrieved, 7,888 after Boolean enforcement, 2,228 after the tier field rule). SpringerLink matched full text rather than metadata, so the Tier B title rule was applied after export (1,694 unique records to 100); Tier A was not executable at that source, because the interface provides no abstract-restricted field. ScienceDirect applied full-text matching under its general query parameter, so the title-abstract-keyword parameter was utilised instead. For Scopus Tier B, the TITLE-ABS-KEY × TITLE-ABS-KEY variant (9,623 hits) was superseded before export by the TITLE-ABS-KEY × TITLE variant, whose 2,266 hits were exported in full; no subset of an adopted result set was exported.

Recall check. A seed set of fifteen papers was assembled before the searches with the intention of spanning the method families; the identified seeds, however, cover only RF, FR, MI and BM, and no seed was recorded for AT, PL, SA or HF, so the check says nothing about search recall in those families. One seed (Paulo et al.) is recorded only by a shorthand that matches four records, and since the intended paper cannot be identified it is removed, leaving 14 seeds. Identification of each seed in the deduplicated records (Appendix D.11) finds 9 unique matches and 1 seed whose shorthand matches several records by the same first author, while 4 seeds have no record: two evaluate GPT-2-scale models only and would be excluded by IC4, and two (Conmy et al. and SAEBench) are search misses. Search recall on the seed set is therefore 10/14, or 10/12 once the two IC4 seeds are set aside. As far as the census is concerned, 8 of the 9 uniquely identified seeds were included, and the remaining one is a survey correctly excluded as a secondary study (X-SEC). Because a seed miss in the chain-of-thought faithfulness literature led to a corrective search pass ("faithful*" treated as direction-locked only when co-occurring with an explanation-context term, adding 208 arXiv, 96 ACL Anthology, 132 Scopus and 6 ACM DL records), the seed set is not an independent validation set, and screening recall is therefore estimated separately on the AI reviewer's exclusions (Section 5.2). The reported Scopus strings were also checked against the exported records, which are consistent with them (Appendix D).

4.4 Eligibility criteria

- **IC1 (publication form):** journal article, conference or workshop paper of at least four pages, or preprint. Publication form is recorded; peer-review status is recorded only where the venue's review process can be verified. The earlier preprint quality gate (citations, public code or "established research affiliation") is withdrawn because it is not an objective quality criterion.
- **IC2 (contribution):** the primary contribution is a method, benchmark, evaluation or empirical study whose object is an explanation or interpretation of LLM behaviour, and the study either evaluates an explanation or makes an explicit explanatory claim about mechanism.

- **IC3 (direction):** the explanandum is a language model or LLM-centred pipeline, named with a quoted Methods-section span.
- **IC4 (model scale and architecture):** at least one explained model is a decoder-only or encoder–decoder language model of ≥1B total parameters; encoder-only models are excluded. IC4 is applied at full text, with one exception at title/abstract: a record is excluded (X-MODEL) when its abstract makes clear that every explained model is encoder-only or clearly small (e.g. GPT-2 small, a toy transformer); whenever model size is the only doubt, the record proceeds to full text.
- **IC5 (date):** first public appearance between 2019-01-01 and 2026-08-28.
- **IC6 (language):** full text available in English.

Exclusion codes are applied in precedence order (X-DUP, X-TYPE, X-SEC, X-DIR, X-TOPIC, X-TRANS, X-MODEL, X-LANG, X-FULL) and the first applicable code is recorded. Secondary studies are diverted to the positioning analysis.

4.5 Deduplication and the superseded automated pilot

Deduplication proceeded by DOI, arXiv identifier, exact normalised title and fuzzy title with first-author agreement, removing 2,095 records (DOI 1,288; exact title 805; fuzzy title 2) and leaving 5,698 unique records; preprint and published versions were merged, retaining the published identifier. The supplementary arXiv search added 32 records with no duplicates, so the census population consists of 5,730 records.

An earlier version of this study screened and coded these records with automated instruments alone, and the resulting pilot (84 machine-coded studies drawn from an availability-conditioned sample) was rightly criticised as unverifiable. It should be noted that the pilot is superseded in its entirety: no pilot decision is used in the census, which re-screened every record from the start, and the pilot is reported in the released material only because its failures motivate the design that follows. In more detail, a plausible-relevance screen disagreed with the strict criteria for 188 of 300 sampled records, a keyword-signal reduction rule recalled only 107 of 725 metadata-eligible records (false-negative rate 85.2%, 95% Wilson CI 82.5–87.6% (Wilson 1927)), a strict metadata screen left the direction rule unenforced (18 X-DIR records in a 120-record audit sample), and five of the eight seed papers that are primary studies in the deduplicated records were excluded by the automation (Appendix D.11). Considering the latter, the census treats every automated decision as a claim that has to carry its own evidence and to be measured against a human reviewer, rather than as a decision that can be trusted by construction.

4.6 Selection and data extraction with a disclosed AI reviewer

The protocol of 2026-09-13 assigned both screening and coding to two human reviewers. Given that a dual human census of 5,730 records, followed by full-text assessment and coding of more than two thousand reports, was not feasible for the author team, Amendment A1 assigns the role of the first reviewer (R1, Coder A) to a disclosed AI reviewer and the role of the second reviewer (R2, Coder B) to a human author (G. Makridis), whose decisions on independent samples are used to calibrate and validate the AI reviewer. The AI reviewer is Claude (model `claude-opus-5-5`, Anthropic),

operated through Claude Code subagents between 2026-10-03 and 2026-10-05. Each batch of records (up to 150 records at title/abstract, about six reports at full text) was processed in a fresh context that received only the frozen instructions, the codebook and the records of the batch, so that the AI reviewer was blind to every decision of the pilot and to every human decision. Every AI decision carries a verbatim quote from the record (title/abstract) or from the report (full text) on which it rests, and a script verifies each quote by string match; decisions whose quote failed were redone until every quote matched. The instructions given to the AI reviewer, all its decisions with their quotes, and all human decisions are released. To this end, the selection proceeded in the following steps:

1. **Calibration.** The AI reviewer and R2 screened the same records independently, with a gate of Cohen's $\kappa \geq 0.70$ (Cohen 1960) on the proceed (include or maybe) versus exclude outcome. When a round failed, the disagreements were reviewed, the instructions were revised as a new version, and a new random set of records was drawn; before the third round it was fixed that no further tuning would follow a third failure.
2. **Screening.** All 5,730 records were screened by the AI reviewer with the instruction version that passed the gate, applying IC4 only under the exception stated above. R2 screened only the calibration and validation samples (450 records); for these records a record proceeds to full text if either reviewer marks it include or maybe, which is how the 41 records sent forward by R2 alone enter the flow, while every other record proceeds or is excluded on the AI reviewer's decision.
3. **Retrieval.** Full texts were retrieved from local copies, arXiv, the ACL Anthology and the best open-access location reported by OpenAlex, and each file was accepted only if the record's title appears in its first two pages; the remaining reports were sought through the University of Piraeus network, by a throttled script that stops at the first refusal and by R2 in a browser, and were subjected to the same identity check. Unretrieved reports are counted by reason, and retrievability-based replacement is prohibited.
4. **Eligibility and coding.** Every retrieved report was assessed against IC1–IC6 and, if eligible, coded in the same pass with the frozen codebook, with quoted evidence for the explanandum, model scale, D1, D3, D4 and E1. Every eligible retrieved study is coded (census); the sampling contingency of the original protocol, which existed to bound human workload, is withdrawn.
5. **Validation.** The AI reviewer's decisions are validated in three ways: (a) the calibration rounds; (b) a simple random sample of 250 of the AI reviewer's exclusions, screened blind by R2, which estimates the rate at which relevant records were wrongly excluded and bounds screening recall; and (c) blind full-text assessment and coding by R2 of a simple random sample of 100 assessed reports, together with every retrieved report that R2 sent forward from the exclusion sample, from which agreement on eligibility and per-field κ are reported. In addition, R2 decides every report the AI reviewer marked unclear. The protocol planned a 20% sample and the verification of every field with $\kappa < 0.60$ for all included studies; for reasons of reviewer time, both were replaced before the sample was drawn, and a field with

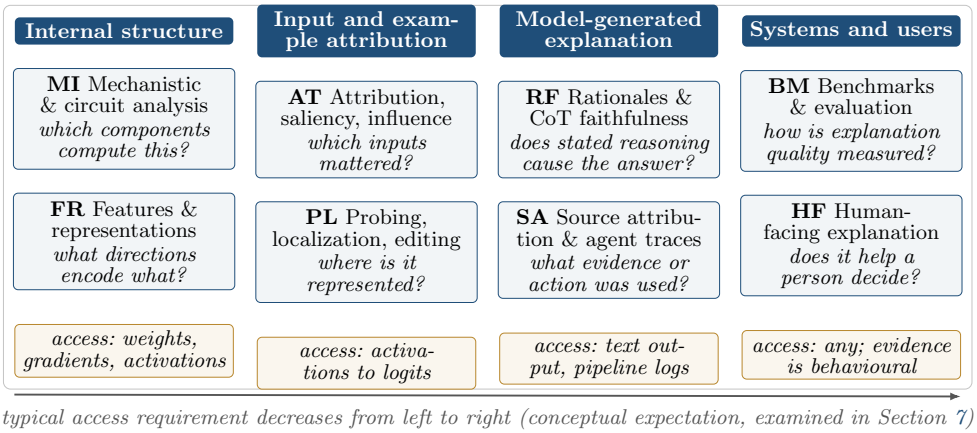


Fig. 1 Conceptual organisation of the method families (D1) by explanandum (columns), with the access regime each family is expected to require a priori. The ordering is a conceptual expectation rather than a coded result; the access regimes actually required by the coded studies are reported in Section 7, and the evidence profile deliberately does not rank interventional against human-subject evidence.

$\kappa < 0.60$ is reported as unreliable and the results that rest on it are flagged instead (Appendix G).

The planned 20% random sample of 1,126 records for R2 at title/abstract was replaced, before any record was drawn, by the exclusion sample of step 5(b), which targets the error that matters for a census (records that are wrongly excluded), whereas records wrongly sent forward are corrected at full text. This and the other deviations are listed in Appendix G, and the instructions, decision files, sampling seeds and analysis scripts are released (Declarations, Data availability).

4.7 Data items and classification framework

Each coded study is classified along four dimensions and an evidence profile (Table 3), and Figure 1 depicts the conceptual organisation of the method families. Definitions, scoring anchors and positive, negative and boundary examples for every code are given in the codebook.

The evidence profile replaces the ordinal D5 scale of the earlier version, in which V3 (controlled intervention) and V4 (human study or deployed system) were ranked on a single axis although they concern different validity dimensions, and in which a study with one or two raters was ranked above a well-controlled causal experiment. In the census, every field is coded from the full text: E1 is I2 if and only if an intervention is compared with a random, shuffled or matched-null control, E2 is decided by the number of human raters or participants alone, E3 records whether the evaluation involves a deployed system, production data or real users, and ROB is restricted to the robustness of the explanation (robustness of model accuracy alone is not ROB). D3 records the access that the method requires, decided from what its computation reads or changes and supported by a quoted span; where no span determines it, the

Table 3 Classification framework. D1 has one primary and optional secondary codes; D4 is multi-coded; the evidence profile fields E1–E4 are independent and not ordered with respect to each other.

Dimension	Code	Definition
D1: Method family	MI	Mechanistic interpretability and circuit analysis
	FR	Feature and representation discovery or interpreted manipulation (SAE, dictionary learning, steering)
	AT	Attribution, saliency and influence
	PL	Probing, localisation and editing with a localisation claim
	RF	Rationales, self-explanation and chain-of-thought faithfulness
	SA	Source attribution (RAG) and tool or agent trace explanation
	BM	Benchmarks and evaluation frameworks for explanations
	HF	Human-facing explanation and interaction
D2: Explanandum	TOK token; SPN span or answer; EX example; NEU neuron or feature; CIR circuit; LAY layer or region; TRJ trajectory; PIPE pipeline	
D3: Access regime (minimum required)	W	Weights and/or gradients of the explained model
	A	Forward activations (read or write), no gradients
	G	Gray-box: training or modifying the model (fine-tuning, adapters, editing)
	L	Output logits or token probabilities only
	T	Text output only
D4: Evaluation	FTH	Faithfulness or causal validity of the explanation (pooled in one code)
	PLA	Plausibility: agreement with human (or declared model-judge) judgement
	HU	Human utility: outcome measured for human participants
	ROB	Robustness or stability
	DWN	Downstream task performance
	NON	None of FTH, PLA and HU; assigned whenever all three are absent (ROB and DWN may still be coded)
E1: Intervention	I0/I1/I2	none / intervention with non-random comparator / intervention with random, shuffled or matched-null control
E2: Human subjects	H0/H1/H2	none / 1–9 / $\geq$10 human raters or participants
E3: Ecological	DEP0/DEP1	laboratory only / deployed system, production data or real users in context
E4: Statistics	S0/S1/S2	neither / one / both of multiple seeds or runs and variance or CIs

code is inferred and marked as such. Additional fields record the largest explained model, total and active parameters from the model card or technical report, availability (open-weight, API-only, not reported), code availability, compute, human-study size and agreement, and regulatory framing (none, sector, GDPR, EU AI Act (European Parliament and Council of the European Union 2024), other).

4.8 Quality appraisal and sensitivity analysis

Clinical risk-of-bias instruments presuppose participants, allocation and blinding and do not transfer to computational method papers; therefore, a twelve-item descriptive rubric adapted from SLR quality-assessment guidance (Kitchenham and Charters 2007) is utilised (Appendix F). Each item is scored 0, 0.5 or 1 against written anchors, Q10 is NA when no human study is reported, and the score is normalised by the number of applicable items; equal weighting is used because no validated weighting exists for these studies, and this is declared as a limitation. The score is never an eligibility filter. It enters the analysis through a sensitivity analysis restricted to studies with normalised score $\geq$0.70, reported in Section 7.7; the raw score restriction of the earlier version is not used, because the denominator varies with Q10. Rubric scores are assigned by the AI reviewer and were not scored by R2, so they rest on the AI reviewer alone.

4.9 Synthesis

The synthesis combines counts and cross-tabulations over the coded fields with a narrative synthesis per method family. All census statistics, validation statistics, tables and figures are generated from the decision files by `census_analysis.py`, which also checks that every flow box reconciles and that every figure draws each study once. The exceptions are the search yields per source (Table 4) and the search record of Appendix D, which are transcribed from the search logs. Beyond its role as reviewer, the same LLM was used under the authors' direction to write the retrieval and analysis scripts and to draft and edit manuscript text, which the authors revised and for which they take full responsibility (see the statement on the use of AI tools). Since the census covers every retrieved and eligible study, counts are reported as "k of n included studies" without sampling intervals; intervals are reported only for the validation samples. Family prevalence is reported for primary codes and for any-code (primary or secondary) counts, because a single exclusive primary code understates families that are frequently secondary contributions.

5 Results: Selection, Validation and Included Studies

5.1 Identification and selection

The searches identified 7,825 records (Table 4), including 32 records from the supplementary arXiv search of 2026-10-03 that covers submissions between 2026-08-19 and 2026-08-28; 2,095 duplicates were removed and 5,730 unique records entered screening. As depicted in Figure 2, 3,315 records were excluded at title/abstract and 2,415 reports were sought for retrieval, of which 2,374 were sent forward by the AI reviewer

Table 4 Records identified and retained by source. The arXiv row includes the supplementary search of 2026-10-03 (214 retrieved, 32 adopted, no duplicates).

Source	Identified	Share	Unique after deduplication
Scopus	3,440	44.0%	3,258
arXiv	2,468	31.5%	1,820
ACL Anthology	795	10.2%	192
IEEE Xplore	612	7.8%	266
ACM Digital Library	269	3.4%	125
ScienceDirect	141	1.8%	53
SpringerLink	100	1.3%	16
Total	7,825	100%	5,730

and 41 only by R2 in the calibration and validation samples. Of the reports sought, 2,212 were retrieved (1,852 through open routes and 360 through institutional access) and assessed at full text, 203 could not be retrieved, 960 were excluded, 0 await R2's decision and 1,252 studies were included. Every box of the flow reconciles arithmetically, and records whose status is unresolved are shown as such instead of being merged with exclusions.

5.2 Validation of the AI reviewer

As far as the calibration is concerned, the AI reviewer did not pass the gate on the first attempt (Table 5). In the first round (100 records), agreement was 79% and $\kappa = 0.54$ (95% CI 0.37–0.71). In 11 of the 21 disagreements the AI reviewer had excluded records whose abstracts could not settle whether an explanation is evaluated, against its own rule that doubt leads to "maybe"; in 3 further disagreements R2 had included studies whose explained models were encoder-only or very small, and the decision was taken to keep IC4 unchanged, since the searches targeted LLM vocabulary only and admitting smaller models would make that part of the map non-systematic. The revised instructions over-corrected: in the second round ($n = 50$) the AI reviewer sent forward 31 records against R2's 19 and $\kappa = 0.32$ (95% CI 0.09–0.56), most disagreements concerning "explainable [domain task] with LLMs" application papers in which an LLM explains its domain predictions and the abstract reports no evaluation of the explanation. Before the third round it was fixed that the round would be the last one and that a failure would switch the design to recall-based validation without further tuning. The third version passed ($n = 50$, agreement 92%, $\kappa = 0.84$, 95% CI 0.67–0.96), and all 5,730 records were re-screened with it.

The calibration measures agreement but not the error that matters most for a census, namely relevant records that are wrongly excluded. To this end, R2 screened blind a simple random sample of 250 of the 3,356 records that the AI reviewer had excluded. R2 would have sent 26 of them to full text, a false-omission rate of 10.4% (Wilson 95% CI 7.2–14.8%), which corresponds to approximately 349 wrongly excluded records.

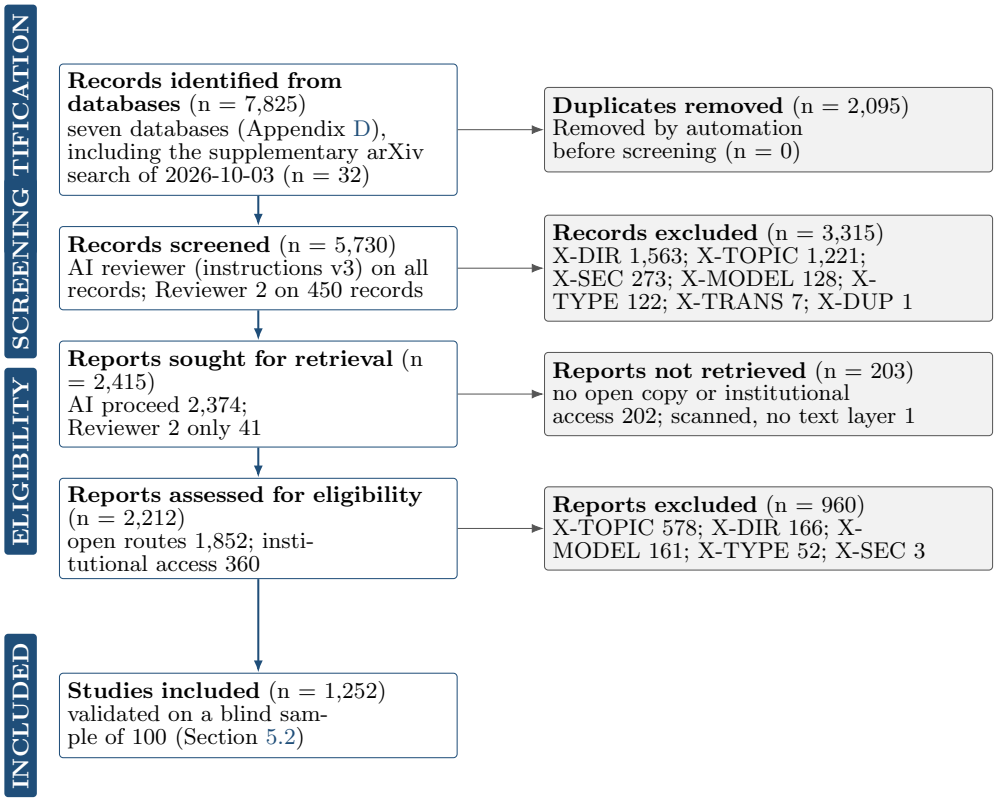


Fig. 2 PRISMA 2020 flow diagram of the census. Every box reconciles: 7,825 − 2,095 = 5,730; 5,730 − 3,315 = 2,415; 2,415 − 203 = 2,212; 2,212 − 960 = 1,252. Screening and full-text decisions were made by the disclosed AI reviewer and validated against Reviewer 2 (Section 5.2).

Under the pre-specified conservative assumption that every such record would be eligible, screening recall is 0.78 (lower bound 0.71), which is below the threshold of 0.95 fixed before the sample was drawn and is reported as a major limitation (Section 9). It is worth mentioning that the assumption is pessimistic: of the 22 of these records that could be retrieved, R2 judged 3 eligible at full text (Wilson 95% CI 5–33%), whereas the AI reviewer had judged none eligible. With R2's eligibility rate, screening recall is 0.96, with a lower bound of 0.88 that remains below 0.95. The misses concentrate in X-TOPIC (21 of 26), which is the boundary between studies that explain an LLM and studies that use explainability as a tool for a task.

As far as the full-text assessment is concerned, R2 assessed blind a simple random sample of 100 retrieved reports (all 100 assessed). On eligibility, agreement was 88% and $\kappa = 0.72$ (95% CI 0.57–0.86; $n = 99$, yes vs no; the remaining 1 sampled report was marked unclear by the AI reviewer and cannot enter a yes/no comparison), with the paired decisions given in Table 6. The disagreements were not symmetric: the AI reviewer included 61 of the 71 sampled reports that R2 judged eligible (sensitivity 0.86, 95% CI 0.76–0.92), while R2 judged 10 of the 36 sampled AI exclusions eligible (28%, 95% CI 16–44%), again mostly at the X-TOPIC boundary. When R2 re-read

Table 5 Calibration of the AI reviewer against Reviewer 2 on independent records (proceed vs exclude). Each round used a new random set of records and the instruction version shown; the gate was Cohen's $\kappa \geq 0.70$. Intervals: bootstrap 95% CI for κ, Wilson 95% CI for recall of Reviewer 2's proceed decisions.

Round	Instructions	n	Agreement	κ (95% CI)	AI recall (95% CI)	Gate
Round 1	v1	100	79%	0.54 (0.37–0.71)	0.64 (0.48–0.77)	failed
Round 2	v2	50	64%	0.32 (0.09–0.56)	0.84 (0.62–0.94)	failed
Round 3	v3	50	92%	0.84 (0.67–0.96)	0.91 (0.72–0.97)	passed

Table 6 Full-text eligibility decisions of the AI reviewer against Reviewer 2's first, blind decisions on the random sample ($n = 99$; the remaining sampled report was marked unclear by the AI reviewer and cannot enter a yes/no comparison).

	R2 eligible	R2 not eligible	Total
AI eligible	61	2	63
AI not eligible	10	26	36
Total	71	28	99

the 12 disagreements against the codebook, 11 were resolved as exclusions and 1 as an inclusion (three of the exclusions only after a second reading), so no adjudication was required; the resolved decisions replace the AI reviewer's in the census, whereas the agreement statistics use R2's first, blind decisions. R2 also decides the reports that the AI reviewer marked unclear, most of which name no explained model or no model size; these are excluded under IC4 (X-MODEL), since eligibility requires the criterion to be shown, and this rule is reported as a clarification of IC4 made during verification. Every report that the AI reviewer marked unclear was decided by R2 from the PDF, so no report remains unresolved; for the reports that R2 included, eligibility and the primary family are R2's, and the AI reviewer coded the remaining fields afterwards with the same instructions. Considering the latter, R2's first decisions imply that a substantial share of the AI reviewer's full-text exclusions may be eligible, whereas the resolved decisions imply far fewer; both readings are reported, and the map should be read as possibly incomplete at this boundary (Section 9).

Per-field agreement over the 61 sampled reports that both reviewers judged eligible is given in Table 7. Method family, explanandum, access regime (in particular the split between internal and output-level access), plausibility and explanation robustness reached $\kappa \geq 0.60$, whereas the intervention type and random control (E1), human-subject evidence (E2), statistical reporting (E4), faithfulness and downstream evaluation did not, and these fields are flagged as unreliable. It should be noted that

Table 7 Agreement between the AI reviewer and Reviewer 2 on the coded fields, over the 61 sampled reports that both judged eligible (R2's first, blind decisions). κ: Cohen's kappa with bootstrap 95% CI. Fields with $\kappa < 0.60$ are flagged as unreliable under the rule fixed before the sample was drawn; for rare codes (HU, NON) κ is low despite high raw agreement because almost all reports fall in one category.

Field	n	Agreement	κ (95% CI)	Flag
D1 primary family	61	80%	0.75 (0.61–0.87)	
D2 explanandum	61	84%	0.80 (0.67–0.90)	
D3 access regime	61	77%	0.61 (0.43–0.77)	
D3 internal (W/A) vs other	61	97%	0.91 (0.76–1.00)	
E1 intervention	61	79%	0.59 (0.38–0.78)	unreliable
E1 random control (I2)	61	84%	0.49 (0.21–0.74)	unreliable
E2 human subjects	61	88%	0.34 (0.00–0.68)	unreliable
E3 deployment	61	100%	–	single category
E4 statistical reporting	61	67%	0.46 (0.28–0.63)	unreliable
Availability	61	93%	0.68 (0.30–0.94)	
D4 FTH	61	80%	0.55 (0.32–0.76)	unreliable
D4 PLA	61	97%	0.90 (0.73–1.00)	
D4 HU	61	95%	0.00 (0.00–0.00)	unreliable
D4 ROB	61	87%	0.63 (0.36–0.85)	
D4 DWN	61	72%	0.26 (0.04–0.49)	unreliable
D4 NON	61	90%	0.00 (0.00–0.00)	unreliable

the direction of the differences is informative: in R2's codes, random controls, human-subject evidence and the reporting of seeds and variance are rarer than in the AI reviewer's codes, while faithfulness evaluation is more frequent, so the corresponding prevalences reported below are more likely to overstate the former and to understate the latter than the reverse. The counts that follow are the AI reviewer's codes, except for the reports decided by R2, and the AI reviewer's own confidence is low for 155 of 1,252 (12.4%) studies, medium for 809 of 1,252 (64.6%) and high for 281 of 1,252 (22.4%).

5.3 Characteristics of the included studies

The 1,252 included studies are concentrated in the most recent years: 594 of 1,252 (47.4%) first appeared in 2026 (to 2026-08-28) and 484 of 1,252 (38.7%) in 2025, against 141 of 1,252 (11.3%) in 2024 and 31 of 1,252 (2.5%) in 2023; only 1 and 1 included studies date from 2022 and 2021 (years are those of first appearance, so 1 journal article assigned to a 2027 volume but published online in 2026 is counted under 2026). Given that the search vocabulary is that of the LLM era (e.g. "large language model", ChatGPT, LLaMA), this distribution partly reflects the search itself, since earlier work on large models often used terms such as "language model" or "GPT-2" (Section 9). As far as the explanandum is concerned (RQ2), neurons and features are explained most often (464 of 1,252 (37.1%)), followed by spans or answers (317 of 1,252 (25.3%)), layers or parameter regions (227 of 1,252 (18.1%)), circuits (104 of 1,252 (8.3%)) and tokens (98 of 1,252 (7.8%)), while training examples (16), agent

Table 8 Profile of the included studies by primary method family. Cells give counts with the within-family percentage in parentheses; *N* gives primary-code / any-code counts, all other cells use the primary code. W/A: weights or activations; G: gray-box; L/T: logits or text only; NON: none of FTH, PLA and HU; I2: intervention with a random, shuffled or matched-null control; Hum.: any human-subject evidence; S2: both seeds and variance reported.

Family	*N*	W/A	G	L/T	FTH	PLA	HU	NON	I2	Hum.	S2
MI	238 / 306	238 (100)	0 (0)	0 (0)	225 (95)	4 (2)	0 (0)	13 (5)	115 (48)	17 (7)	58 (24)
FR	363 / 458	358 (99)	1 (0)	4 (1)	289 (80)	77 (21)	2 (1)	54 (15)	90 (25)	40 (11)	69 (19)
AT	137 / 169	59 (43)	1 (1)	77 (56)	87 (64)	51 (37)	11 (8)	24 (18)	24 (18)	34 (25)	30 (22)
PL	171 / 255	168 (98)	3 (2)	0 (0)	104 (61)	7 (4)	0 (0)	63 (37)	38 (22)	20 (12)	36 (21)
RF	238 / 281	20 (8)	78 (33)	140 (59)	133 (56)	139 (58)	9 (4)	5 (2)	30 (13)	105 (44)	44 (18)
SA	31 / 38	2 (6)	3 (10)	26 (84)	7 (23)	26 (84)	5 (16)	0 (0)	0 (0)	16 (52)	3 (10)
BM	55 / 115	26 (47)	0 (0)	29 (53)	35 (64)	26 (47)	0 (0)	7 (13)	15 (27)	14 (25)	12 (22)
HF	19 / 44	3 (16)	0 (0)	16 (84)	1 (5)	5 (26)	15 (79)	2 (11)	0 (0)	17 (89)	5 (26)

trajectories (9) and full pipelines (17) are rare. Online Resource 1 lists every included study with all its codes and the quotes that support them.

6 Method Families

This section describes each method family conceptually and through representative coded studies, while Table 8 reports the quantitative profile of each family; the numbers are not repeated in the text. The studies cited in this section are representative examples; every included study is listed, with its codes and supporting quotes, in Online Resource 1.

6.1 Mechanistic interpretability and circuit analysis (MI)

Mechanistic interpretability seeks explanations in internal computation (i.e. circuits, attention heads, residual-stream directions and causal dependencies), with an interventional evidence standard: if a component is patched, ablated or steered, behaviour should change as predicted. The synthesis must distinguish descriptive localisation from causal explanation, since a layer or direction correlated with a behaviour does not establish that it causes the behaviour, and a successful intervention demonstrates local causal relevance without explaining the whole behaviour. Representative coded studies analyse arithmetic in-context learning circuits in Pythia-12B (Polyakov et al. 2025), positional bias in multiple-choice answering (Li and Gao 2025), the role of latent romanization in multilingual processing (Saji et al. 2025), reasoning-to-answer information flow in reasoning models (Zhang et al. 2025), and where chain-of-thought reasoning commits to an answer (Zhang et al. 2026). In the census, the family is almost entirely W/A, about half of its studies compare an intervention with a random or matched-null control, and human-subject evidence is rare. The limitations recorded by these studies concern restriction to single model families or narrow diagnostic tasks, incomplete circuit identification, and automated rather than human validation of feature interpretations.

6.2 Feature discovery and representation methods (FR)

This family includes sparse autoencoders, dictionary learning, representation engineering, and steering or concept vectors, i.e. methods that discover or manipulate interpretable directions in activation space. They sit between explanation and control: they can reveal recurring functional structure, but they require validation to avoid mistaking decomposability for faithful explanation. Representative studies include contrastive activation addition for steering Llama 2 (Rimsky et al. 2024), feature extraction at scale for Llama-3.1-8B (He et al. 2024), low-rank adaptation of models for SAE fidelity (Chen et al. 2025b), and the frame representation hypothesis for multi-token concepts (Valois et al. 2025). It is worth mentioning that FR is the largest family of the census, both as primary code and with secondary codes, since feature methods are also a frequent secondary contribution of MI and PL studies; studies that utilize SAE features or steering only to improve a task were excluded as X-TOPIC at full text. Recorded limitations include sensitivity to layer, norm and steering-strength choices, capability trade-offs under steering, and manual feature selection.

6.3 Attribution, saliency and influence (AT)

Attribution assigns importance to input tokens, generated tokens, training or demonstration examples, or interactions among inputs, and is therefore attractive for local explanation but vulnerable to instability, baseline dependence and the conflation of faithfulness with plausibility. Token attribution, example influence and source attribution are distinct questions that require different evidence. Representative studies include scalable feature-interaction attribution with logit access (Kang et al. 2025), ridge-regression influence estimation (Tu et al. 2025), multi-stage influence functions (Bao et al. 2025), and query-relevant neuron attribution (Chen et al. 2025a). Access in AT is divided almost evenly between internal access and logits or text, and several applied studies explain an LLM classifier with LIME or SHAP, whether the model is used as released (Kastrati et al. 2025) or fine-tuned for the task (Lodh et al. 2025); the codebook treats both as boundary cases for IC2.

6.4 Probing, localisation and editing (PL)

Probing and localisation ask where information or behaviour is represented, whereas editing tests whether a targeted modification changes behaviour. The key synthesis question is whether localisation evidence is predictive or causal, and a second question is whether edits preserve unrelated capabilities, since explanation and intervention become entangled. Representative studies identify linear relational concepts (Chanin et al. 2024), localise sycophantic agreement (Duszenko et al. 2026), show that linear probes can detect task format rather than reasoning mode (Sahoo et al. 2026), and attribute query-relevant knowledge to specific neurons (Chen et al. 2025a). Knowledge-editing studies are included only when they make or test a localisation claim; editing studies evaluated only on editing reliability and downstream performance were excluded as X-TOPIC, which is the codebook's boundary for this family.

6.5 Rationales, self-explanation and chain-of-thought faithfulness (RF)

This family examines explanations that the model produces in natural language, and asks whether stated reasoning reflects the process that caused the answer. Plausibility and faithfulness must be kept separate, since human agreement with a rationale is not evidence that the rationale caused the answer (Jacovi and Goldberg 2020). Representative studies demonstrate that chain-of-thought explanations can systematically misrepresent the reasons for a prediction (Turpin et al. 2023), train models for counterfactual simulatability of their reasoning (Hase and Potts 2026), utilize causal importance to improve faithfulness (Swaroop et al. 2025), and show that interpretable reasoning traces can diverge from final-answer quality in a human study (Bhambri et al. 2025). Given that RF methods mostly operate on text, the family holds the largest share of the API-compatible access in the census, namely 140 of the 292 studies coded L or T.

6.6 Source attribution and agent-trace explanation (SA)

Retrieval augmentation raises the question of which retrieved documents support an answer, and tool-using agents add trajectories of plans, calls and observations; answer attribution and process attribution are different claims. Only 31 included studies carry SA as their primary family and 38 with any code, among them a mechanistic study of how LLMs cite their sources (van Dort and Heuss 2026). Most of the family operates on text and evaluates attributions against human judgements, so plausibility dominates and faithfulness is evaluated in about a quarter of its studies (Figure 8). It should be noted that agent and retrieval systems whose traces are not evaluated as explanations were excluded as X-TOPIC, and that the small size of the family may reflect the search vocabulary (Section 9) as much as the literature.

6.7 Benchmarks and evaluation frameworks (BM)

Benchmarks define what the field can compare. Representative studies evaluate sparse autoencoders at scale (Gao et al. 2025), test whether global XAI methods recover behaviours injected into an LLM (Sovrano 2025), and show that measured chain-of-thought unfaithfulness can be disguised accuracy (Bentham et al. 2024). The any-code count of BM (115) is twice its primary count (55), since benchmark construction is often a secondary contribution of method papers, whereas task benchmarks without an evaluation of explanations were excluded as X-TOPIC.

6.8 Human-facing explanation and interaction (HF)

Human studies test whether explanations improve understanding, calibration, trust, debugging or task performance, and a method can be useful to a user while failing as a causal explanation. Only 19 included studies carry HF as their primary family, and 44 with any code, among them a human study of interpretable reasoning traces (Bhambri et al. 2025) and an interpretable and editable decision-making framework evaluated with experts (He et al. 2026). Human-facing work is thus present mainly

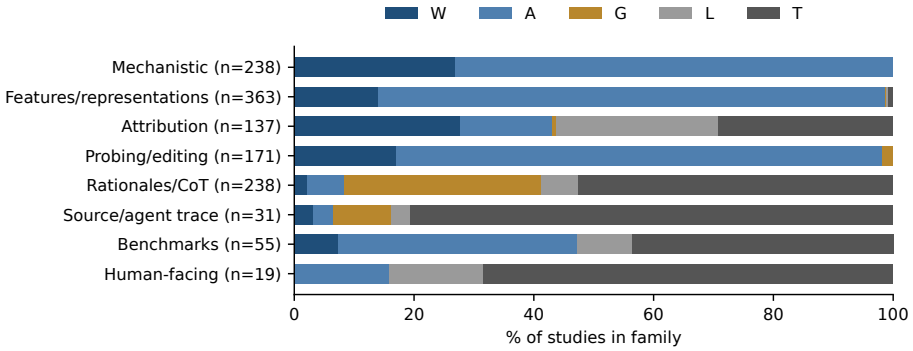


Fig. 3 Access regime required by the explanation method (D3), as a share of the included studies of each primary method family; n is the number of included studies in the family (1,252 in total). W: weights or gradients; A: forward activations; G: training or modifying the model; L: logits only; T: text output only.

as a secondary element of RF and AT studies rather than as a primary empirical programme, which is consistent with the rarity of human-utility evaluation across the census (Section 7).

7 Cross-Cutting Analysis

7.1 Access regime (RQ3)

As far as the access regime is concerned, methods requiring internal access dominate the census: W (weights and/or gradients) is required in 192 of 1,252 (15.3%) studies and A (activations) in 682 of 1,252 (54.5%), giving 874 of 1,252 (69.8%) studies that require W or A. Gray-box methods (G), which train or modify the model through fine-tuning, adapters or editing but do not inspect its internals, account for 86 of 1,252 (6.9%), while logit-only access (L) is required in 62 of 1,252 (5.0%) and text-only access (T) in 230 of 1,252 (18.4%). Given that W and A are unavailable through ordinary hosted APIs, that L is available only where a provider exposes token probabilities, and that G is available only through the fine-tuning endpoints that some providers offer, the access regimes compatible with hosted models (L and T) account for 292 of 1,252 (23.3%) studies. As depicted in Figure 3, the split follows the families closely: MI is entirely W/A (238 of 238 studies) and FR and PL almost entirely so, whereas RF and SA mostly operate on text, AT and BM are divided between internal and output-level access, and RF is the only family in which gray-box methods (i.e. training or fine-tuning the model) are common. The code is inferred rather than supported by a quoted span in 23 studies. It should be noted that D3 is a single code and the codebook gives no explicit rule for a method that needs several resources (e.g. a method that fine-tunes a model and then reads its activations); such cases were decided by the AI reviewer from the computation the method describes. For reading the results, the code is interpreted as the most internal resource the method needs, in the order W, A, G, L, T. The split between internal access (W/A) and the other regimes, on which

Table 9 Scale and availability of the largest explained model per study (median reported size 8.0B).

Characteristic	Studies
Parameter count reported or documented	1,108 of 1,252 (88.5%)
of which ≥70B	165 of 1,252 (13.2%)
of which ≤10B	712 of 1,252 (56.9%)
Undisclosed (proprietary) or not stated	144 of 1,252 (11.5%)
Explained model open-weight	1,107 of 1,252 (88.4%)
Explained model API-only	133 of 1,252 (10.6%)
Availability not reported	12 of 1,252 (1.0%)

the main result rests, reached $\kappa = 0.91$ against R2 (Table 7), but a method that combines training access with internal access may have been coded either way, which is a limitation of the G count in particular.

7.2 Model scale and availability

Parameter counts are recorded only from the paper, the model card or the technical report (Table 9), and proprietary models without an official parameter count (e.g. GPT-4 (OpenAI 2023), GPT-4o) are recorded as undisclosed rather than estimated. Among the 1,108 studies with a documented count, the median largest explained model has 8.0B parameters; 712 of 1,252 (56.9%) studies explain no model larger than 10B, and 165 of 1,252 (13.2%) explain a model of at least 70B. It is evident that the census is dominated by open-weight models (1,107 of 1,252 (88.4%)), whereas API-only models are the largest explained model in 133 of 1,252 (10.6%) studies, which is consistent with the access regimes above: the models that most users interact with are the ones that the prevalent methods cannot open.

7.3 Faithfulness, plausibility and human utility (RQ4)

Faithfulness or causal-validity evidence (FTH, which pools tests that an explanation reflects the computation with interventional tests of causal relevance) was coded in 881 of 1,252 (70.4%) studies; this is an AI-coded estimate with $\kappa = 0.55$ against R2 and should be read as approximate. Plausibility was evaluated in 335 of 1,252 (26.8%), of which 147 evaluated both, while human utility (i.e. an outcome measured for human participants) was evaluated in only 42 of 1,252 (3.4%). Downstream performance (762 of 1,252 (60.9%)) and the robustness of the explanation (209 of 1,252 (16.7%)) are also frequent. The 168 studies with none of FTH, PLA and HU are included through an explicit explanatory claim about mechanism or through an evaluation of explanation robustness, and 122 of them carry no evaluation of the explanation at all; these are the studies whose eligibility under IC2 rests most on judgement. As depicted in Figure 4, faithfulness dominates in MI, FR and PL, plausibility exceeds faithfulness by a wide margin in SA, and RF evaluates both in comparable shares. Human utility is most

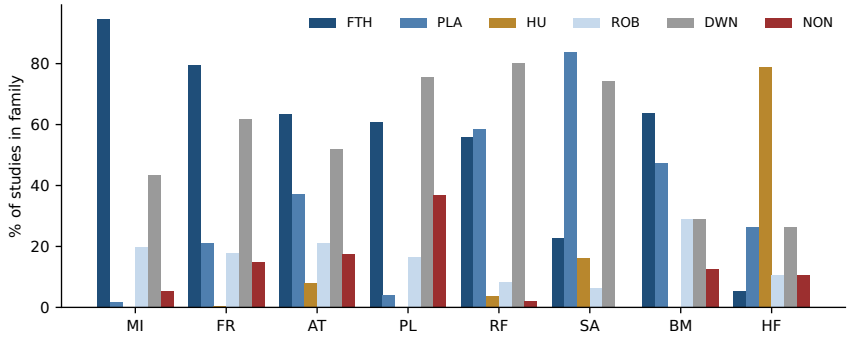


Fig. 4 Evaluation constructs (D4) by primary method family, as a share of the included studies of each family. D4 is multi-coded, so the shares of a family can sum to more than 100%; NON marks studies with none of FTH, PLA and HU.

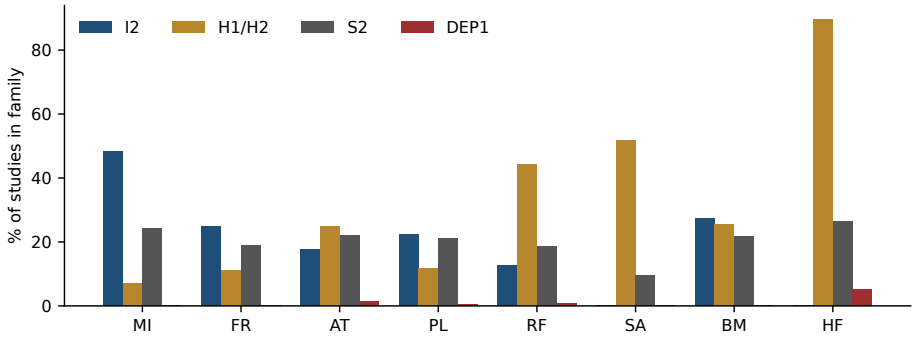


Fig. 5 Evidence profile by primary method family: share of studies with a random-controlled intervention (I2), any human-subject evidence (H1/H2), both seeds and variance reported (S2) and a deployed or real-user setting (DEP1). The fields are independent and not ordered with respect to each other.

concentrated in HF, where it is the main construct (15 of 19 studies), but it is also coded in AT (11), RF (9), SA (5) and FR (2).

7.4 Evidence profile (RQ5)

Interventional evidence (I1 or I2) was coded for 1,055 of 1,252 (84.3%) studies, but only 312 of 1,252 (24.9%) compared the intervention with a random, shuffled or matched-null control (Table 10). Any human-subject evidence appeared in 263 of 1,252 (21.0%) studies, and in 187 of these fewer than ten people were involved, while a deployed system, production data or real users in context (E3) appear in only 7 studies. Moreover, both multiple seeds and variance were reported in 257 of 1,252 (20.5%). It is obvious from Figure 5 that random-controlled interventions concentrate in MI, whereas human-subject evidence concentrates in HF, SA and RF; these are different kinds of evidence, and the study makes no claim that one is stronger than the other. It

Table 10 Evidence profile of the included studies. The four fields are independent and not ordered with respect to each other.

Field	Level	Studies
Intervention (E1)	I0 no intervention	197 of 1,252 (15.7%)
	I1 intervention, no random, shuffled or matched-null control	743 of 1,252 (59.3%)
	I2 intervention with random, shuffled or matched-null control	312 of 1,252 (24.9%)
Human subjects (E2)	H0 none	989 of 1,252 (79.0%)
	H1 1–9 human raters or participants	187 of 1,252 (14.9%)
	H2 10 or more	76 of 1,252 (6.1%)
Ecological (E3)	DEP0 laboratory only	1,245 of 1,252 (99.4%)
	DEP1 deployed system or real users	7 of 1,252 (0.6%)
Statistical (E4)	S0 neither seeds nor variance	559 of 1,252 (44.6%)
	S1 one of the two	436 of 1,252 (34.8%)
	S2 both	257 of 1,252 (20.5%)

should be noted that interventional evidence establishes that an explanation predicts the effect of an intervention, whereas a human-rating study establishes agreement or utility for people, and neither establishes the other.

7.5 Reporting indicators

Multiple seeds or runs were reported in 329 of 1,252 (26.3%) studies, variance, confidence intervals or significance tests in 621 of 1,252 (49.6%), a link to code in 625 of 1,252 (49.9%) and compute in 345 of 1,252 (27.6%). It should be noted that the code indicator records whether the paper gives a link, not whether the link resolves to working code, and resolution was not checked. The 7 studies included by R2 against the AI reviewer's decision were coded by R2 on the fields of Table 3 only, so they carry no code for code link, compute or regulatory framing and are counted as not reporting them.

7.6 Regulatory framing (RQ6)

Only 10 of 1,252 (0.8%) studies connect their explanation outputs to a specific regulatory or sector obligation (EU AI Act (European Parliament and Council of the European Union 2024) 5, GDPR 1, sector-specific guidance 4), while 1,235 studies

Table 11 Quality sensitivity analysis: all included studies vs studies with normalised quality score ≥0.70. Q1 and Q9 reward an explicit explanation claim and a faithfulness or plausibility evaluation, so differences in the construct rows are expected by construction.

Indicator	All included	Quality ≥0.70
Weights or activations required (W/A)	874 of 1,252 (69.8%)	373 of 505 (73.9%)
Logits or text only (L/T)	292 of 1,252 (23.3%)	106 of 505 (21.0%)
Faithfulness evaluated (FTH)	881 of 1,252 (70.4%)	422 of 505 (83.6%)
Human utility evaluated (HU)	42 of 1,252 (3.4%)	14 of 505 (2.8%)
None of FTH, PLA and HU (NON)	168 of 1,252 (13.4%)	37 of 505 (7.3%)
Random-controlled intervention (I2)	312 of 1,252 (24.9%)	208 of 505 (41.2%)
Any human-subject evidence (H1 or H2)	263 of 1,252 (21.0%)	118 of 505 (23.4%)
Seeds and variance (S2)	257 of 1,252 (20.5%)	187 of 505 (37.0%)

make no such connection and 7 studies coded by R2 are not coded for this field. Considering the latter, the frequent appeal to transparency obligations in the motivation of explanation research is rarely carried through to the level of a stated requirement that a specific explanation property is meant to satisfy.

7.7 Quality sensitivity analysis

The median normalised quality score was 0.7 and 505 studies scored ≥0.70 (Table 11). Restricting to these studies raises the share of studies with a faithfulness evaluation and lowers the share of NON studies; this is expected rather than discovered, because Q1 rewards an explicit explanation claim and Q9 rewards a faithfulness or plausibility evaluation, so the score partly encodes the properties that define NON, and both the score and the codes come from the same AI reviewer. The access imbalance and the rarity of human-utility and human-subject evidence persist under the restriction, which is the result that the sensitivity analysis is meant to test. Figure 6 depicts the growth of the literature by method family, and Figure 7 the access regime by year.

8 Gaps and Research Directions

Figure 8 cross-tabulates primary family against eight dimensions of the review. Taking into account that the codes are those of the AI reviewer and that several evidence fields are unreliable (Table 7), four gaps are consistent across the families and the

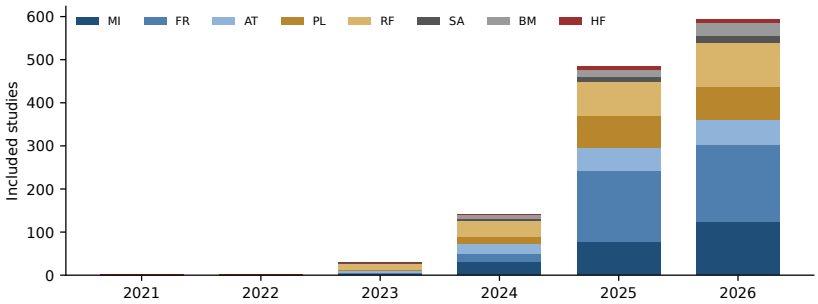


Fig. 6 Included studies per year of first appearance, by primary method family (2026 covers 2026-01-01 to 2026-08-28).

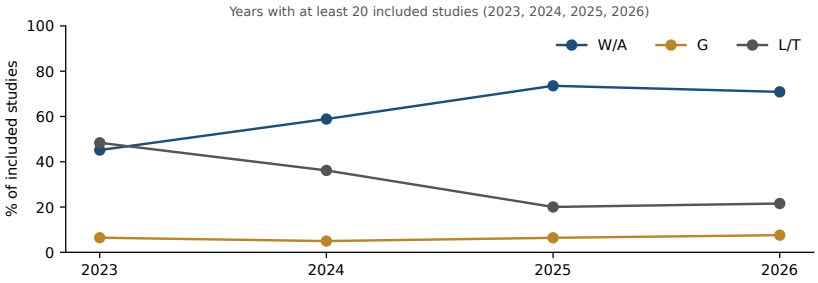


Fig. 7 Share of included studies requiring weights or activations (W/A), training access (G) or logits or text only (L/T), for each year with at least 20 included studies.

sensitivity analysis, and for the gaps that rest on unreliable fields (human-subject evidence, controls and statistical reporting) the human reviewer's codes point in the same direction, and they motivate the following research directions:

1. **Explanation methods compatible with hosted access.** MI contains 0 and FR 4 studies that operate on logits or text. Research on behavioural or logit-level proxies for mechanistic findings, and on validating such proxies against white-box ground truth on open-weight models, would allow explanations to be applied where most LLMs are deployed.
2. **Human-subject validation.** Human-utility evaluation appears in 42 of 1,252 (3.4%) studies, and most human evaluations involve fewer than ten people. Controlled user studies that measure understanding, calibration or debugging outcomes, reported with agreement statistics, are needed before explanations are presented to users, auditors or domain experts as fit for purpose.
3. **Control adequacy and statistical reporting.** Fewer than a third of the interventional studies compared the intervention with a random or matched-null control, and 257 of 1,252 (20.5%) studies reported both seeds and variance. Standardised null controls and uncertainty reporting for interpretability interventions would make results comparable across methods.

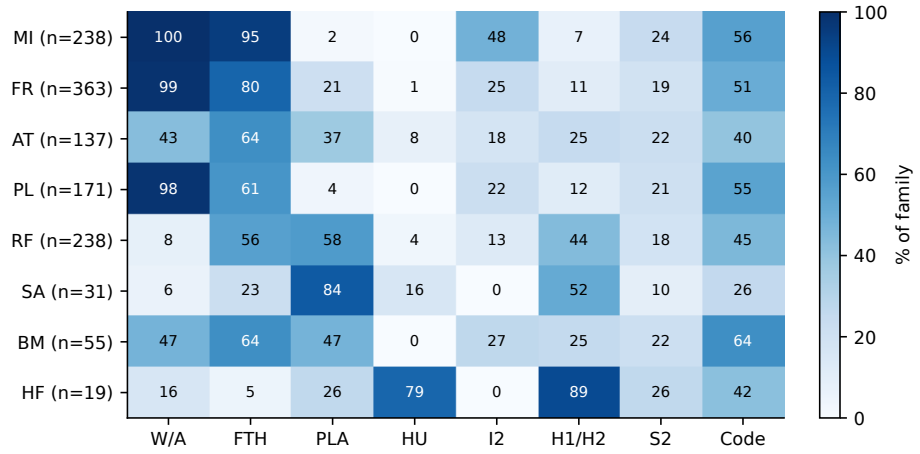


Fig. 8 Coverage matrix: primary method family against review dimension, as the percentage of the included studies of each family (1,252 in total). W/A: weights or activations required; FTH, PLA, HU: evaluation constructs; I2: random-controlled intervention; H1/H2: any human-subject evidence; S2: seeds and variance; Code: a code link is given.

4. **Mapping explanations to obligations.** Only 5 studies mapped their explanation outputs to a specific EU AI Act obligation. Work that states which explanation property (e.g. faithfulness of a rationale, attribution to a retrieved source) satisfies which transparency or oversight requirement would connect the technical literature to its most frequently cited motivation.

In addition, the thin representation of source attribution, agent-trace explanation and human-facing explanation as primary families (31 and 19 studies respectively) is noted; it may partly reflect the search vocabulary and the single-primary-code design, and it is reported as an observation rather than as a gap.

9 Limitations

1. **Screening recall.** Screening was performed by the AI reviewer, and in a random sample of its exclusions R2 would have sent 10.4% of the records to full text (95% CI 7.2–14.8%). Under the pre-specified conservative assumption that all such records are eligible, screening recall is 0.78 (lower bound 0.71), below the threshold of 0.95 fixed in advance. With R2's full-text eligibility rate for the retrieved records of that sample, recall is 0.96, but its lower bound (0.88) is also below 0.95. At full text, the AI reviewer included 0.86 (95% CI 0.76–0.92) of the sampled reports that R2 judged eligible on first reading, although most of these disagreements were resolved as exclusions on re-reading. Readers should therefore treat the map as possibly missing studies at the boundary between explaining an LLM and using explainability as a tool for a task, where the misses at both stages concentrate.
2. **One human reviewer and an AI reviewer.** All screening, eligibility and coding decisions outside the validation samples were made by the AI reviewer, and only

one human reviewer was available; the dual human census of the original protocol was not performed. The calibration needed three rounds, and the gate was passed only by the third instruction version, on 50 records. The full-text validation rests on a blind sample of 100 reports instead of the planned 20%, so its intervals are wide, and fields with $\kappa < 0.60$ (E1, E2, E4, faithfulness and downstream evaluation) are flagged rather than verified for every study.

3. **Unretrieved reports.** 203 of the 2,415 reports sought could not be obtained through open routes or the institutional subscriptions of the University of Piraeus. They are predominantly records without a DOI, reports of smaller publishers and Springer volumes outside the subscription, so the map may under-represent workshop and regional venues.
4. **Search vocabulary and dates.** The searches use LLM-era vocabulary, so earlier work on models of ≥1B parameters that describes them as "language models" (2019–2022) is under-retrieved, and terms for source attribution, citation grounding and agent-trace explanation may be under-represented. Each source covers the records that it had indexed on its search date (Appendix D); arXiv, the largest source, was searched again so that it covers submissions up to the end date of IC5. Recall on the recorded seed set could be confirmed only partially, the seeds cover four of the eight method families, and citation searching was not performed.
5. **Scope.** The study covers the scaling of interpretability methods to models of ≥1B parameters, not the origins of the field; foundational mechanistic work on smaller models is out of scope by IC4.
6. **Coding.** The quality rubric is adapted, equally weighted and not validated. Access regime is coded as the minimum access the method requires; in 23 studies no passage determines it and the code is inferred. Model sizes not stated in a paper were taken from the provider's documentation and were not verified individually.

10 Conclusions

This study presents and defines the problem of mapping how LLMs are explained, and it maps 1,252 studies along an explicit explanandum boundary, an access-regime dimension and a non-ordinal evidence profile, with a selection trail in which every AI decision carries its evidence and every validation step, including the failed calibration rounds, is reported. The map indicates that the explanation methods of the field mostly require internal access (69.8% of the included studies require weights or activations) that the hosted LLMs most users interact with do not expose, that faithfulness or causal-validity evidence is frequently coded (70.4%, an approximate AI-coded estimate) while human utility rarely is (3.4%), that fewer than a third of the interventional studies (312 of 1,055) compare against a random or matched-null control, and that the connection to specific regulatory obligations is almost absent. It should be noted that the study also offers a methodological lesson from this case: in this review, evidence-linked AI decisions and independent human validation on separate samples made the AI reviewer's limitations visible (two failed calibration rounds, a screening recall below the target set in advance and a full-text strictness at the

boundary of IC2) and allowed them to be reported rather than tuned away. Whether the same safeguards are sufficient, or necessary, in other reviews is not tested here.

Future work will be focused around the following directions:

1. behavioural and logit-level proxies for mechanistic explanations, validated against white-box ground truth on open-weight models, so that explanations become applicable where LLMs are deployed;
2. controlled human-subject studies of explanation utility, reported with agreement statistics;
3. standardised null controls and uncertainty reporting for interpretability interventions;
4. an update of the map with a second human reviewer and a search vocabulary that also covers pre-LLM terminology and source-attribution and agent-trace terms.

Supplementary information. Online Resource 1 (`ESM_1.xlsx`) is the master classification file: every included study with all its codes, the model it explains, its quality score, the source of each decision and the verbatim quotes that support the codes, together with every screened record and its decisions. Online Resource 2 (`ESM_2.pdf`) is the codebook with its definitions, boundary rules, quality rubric and amendment log.

Acknowledgements. The research leading to the results presented in this paper has received funding from the European Union's funded Project EXTRA-BRAIN under grant agreement no 101135809. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Declarations

Funding. This research has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101135809 (EXTRA-BRAIN).

Competing interests. G. Makridis is also affiliated with ExpertAI SARL, Luxembourg. The authors have no other relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate. Not applicable: the study analyses published literature, and the only human reviewer is one of the authors.

Consent for publication. Not applicable.

Data availability. The complete review record is deposited on Zenodo (Makridis et al. 2026) (https://doi.org/10.5281/zenodo.23404135). Because Scopus and IEEE Xplore licences restrict redistribution of raw exports, the deposit contains derived metadata only (record identifier, DOI or arXiv identifier, title, year, source, decision and code) rather than raw export files or abstracts, and it contains no full texts. It includes the full search record with superseded strategies (Appendix D); the protocol, Amendment A1 with its dated calibration and validation log, and the codebook; the

frozen instructions given to the AI reviewer (title/abstract versions 1 to 3 and full-text version 1); every AI decision with its supporting quote and every human decision; the calibration, exclusion-sample and full-text verification samples with their seeds; the codes of every included study (also Online Resource 1); the screening desk used by R2; and the scripts that batch the records, check every quote, compute the agreement and recall statistics and regenerate every table, figure and number reported (`census_analysis.py`). The superseded pilot wave and its decision files, and the two discarded automated fills of R2's workbook with their scripts, are included for transparency and are not used in any analysis. Data and documentation are released under CC BY 4.0 and code under the MIT licence.

Code availability. All scripts that batch the records, check every quote, compute the agreement and recall statistics and regenerate every table, figure and number reported are included in the review record above.

Use of generative AI. An AI reviewer is part of the method and is described in Section 4.6: Claude (Anthropic, model `claude-opus-5-5`), operated through Claude Code subagents from 2026-10-03 to 2026-10-05, screened all records, assessed every retrieved report and coded every included study, with frozen instructions that are released and with a verbatim quote for every decision; its decisions are validated against a human reviewer as reported in Section 5.2. The same model was utilised under the authors' direction to write the analysis and retrieval scripts, to build the screening desk used by the human reviewer, and to draft and edit manuscript text. In the superseded pilot wave, automated assistance was provided by OpenAI Codex with GPT-5.5 between 2026-09-02 and 2026-09-12, and Claude (claude-opus-5) was utilised on 2026-09-13 and 2026-09-14 for the revision scripts; no pilot decision is used in the census. Codex with GPT-5.5 was also used on 2026-10-07 to produce two automated fills of the human reviewer's full-text workbook; both were discarded, are released for transparency and are not used in any analysis (Appendix G). The authors take full responsibility for the content.

Author contributions. Conceptualization, G.M., G.F. and J.S.; methodology, G.M. and G.F.; software, G.M.; validation, G.M.; formal analysis, G.M.; investigation, G.M. and G.F.; data curation, G.M.; writing (original draft preparation), G.M. and G.F.; writing (review and editing), G.M., G.F., S.K., J.S. and D.K.; visualization, G.M. and G.F.; supervision, J.S. and D.K.; project administration, G.M. All authors have read and agreed to the published version of the manuscript.

Appendix A Sources for the Positioning Table

Each cell of Table 1 was checked against the cited paper (the version listed in the references). The locations below support the cells; "none" means that the paper was searched for the relevant terms (e.g. search, database, query, PRISMA, inclusion, table, supplementary, repository, white-box, black-box, API, access) and contains no such statement.

Review	Locations
Luo and Specia (2024)	Method: p. 1, Introduction, and p. 2, Section 2 (no search or selection process). Search window: none. Matrix: none (p. 3, Figure 1 is a taxonomy). Access: none ("black-box" denotes opacity, p. 1). Evidence coding: none (p. 8–9, Section 5 discusses evaluation in general).
Zhao et al. (2024)	Method: p. 1, Introduction, and p. 8, Conclusions (described as a systematic overview; no search or selection process). Search window: none. Matrix: none (no tables). Access: p. 13, Appendix A.1. Evidence coding: none.
Cambria et al. (2024)	Method: p. 3–5, Section III (search, selection and Figure 1). Search window: p. 4, Section III-A and III-C (arXiv 2010 to October 2023; 1,030, 233 and 35 records at successive stages). Matrix and access: p. 6, Table I (one row per application paper, with a model-agnostic column; discussion papers in prose, Section IV-B). Evidence coding: none (Table I codes target, agnosticism, goal and repository engagement).
Herrera (2025)	Method: p. 4, Section 1, and p. 26, Section 8 (conceptual and normative reflection). Search window: none. Matrix: none (p. 15, Table 1 lists audiences, not studies). Access: implicit only (p. 8, Section 4; p. 18, Section 6). Evidence coding: none (p. 21, Section 7).
Palikhe et al. (2025)	Method: p. 1, Introduction (described as a systematic review; no search or selection process). Search window: none. Matrix and access: p. 3, Table 1 (22 studies, eight dimensions including an internal/external access column; defined p. 4, Section 4). Evidence coding: none.
Gao et al. (2026)	Method: p. 1, Abstract, and p. 2, Section 1 (described as a systematic review; no search or selection process). Search window: p. 9, Limitations ("up to late 2025"). Matrix: p. 1, Abstract, and p. 9, Section 6 (paper list in a repository). Access: not applicable (p. 2, Section 2.1). Evidence coding: none (p. 8, Section 5).

Appendix B PRISMA 2020 Checklist

Locations are given as section, table, figure and appendix references.

Item	Checklist topic	Location or adaptation
1	Title	Title identifies the report as a systematic mapping study.
2	Abstract	Structured abstract.
3	Rationale	Section 2.
4	Objectives	Section 3.

Item	Checklist topic	Location or adaptation
5	Eligibility criteria	Section 4, Eligibility criteria; codebook Section 1.
6	Information sources	Section 4, Information sources; Appendix D with dates.
7	Search strategy	Appendix D: executed strings and local rules per source, filters and limits.
8	Selection process	Section 4.6: AI reviewer (R1) and human reviewer (R2), independence, frozen instruction versions, three calibration rounds; Section 5.2: agreement and recall.
9	Data collection process	Section 4.6: AI reviewer codes every included study with quoted evidence; R2 assesses and codes a blind random sample of 100 reports; unreliable fields flagged.
10	Data items	Table 3; codebook Sections 2–7.
11	Study risk of bias	Adapted quality rubric with anchors (Appendix F).
12	Effect measures	Not applicable (mapping study).
13	Synthesis methods	Section 4, Synthesis; primary and any-code prevalence; sensitivity analyses.
14	Reporting bias assessment	Section 9, items 1, 3 and 4.
15	Certainty assessment	Not applicable (mapping study).
16	Study selection	Figure 2; Appendix E.
17	Study characteristics	Section 5; Appendix C; Online Resource 1.
18	Risk of bias in studies	Section 7.7; quality scores in Online Resource 1.
19	Results of individual studies	Online Resource 1 (codes and supporting quotes for every included study).
20	Results of syntheses	Sections 6–8.
21	Reporting biases	Limitations.
22	Certainty of evidence	Not applicable (mapping study).
23	Discussion	Sections 8–10.
24	Registration and protocol	Not registered (Section 4); protocols released with the study; deviations in Appendix G.
25	Support	Declarations (Funding).
26	Competing interests	Declarations (Competing interests).
27	Availability of data, code and materials	Declarations (Data and Code availability); Online Resources 1 and 2.

Appendix C Included Studies

With 1,252 included studies, the per-study classification is released as Online Resource 1 (sheet Included_studies) rather than printed. For every included study, it gives the record identifier, DOI or arXiv identifier, title, year and source, the codes D1 (primary and secondary), D2, D3, D4 and E1–E4, the largest explained model with its total parameters and availability, the normalised quality score, the AI reviewer's confidence, the verbatim quotes supporting D1, D3, D4, E1, E2 and E3, and, for the records in R2's samples, R2's values. Every table and figure of Sections 5–8 is regenerated from these data by `census_analysis.py`.

Appendix D Search Record

D.1 Term blocks

Population block (P), used in all Tier A and Tier B strategies:

```
"large language model*" OR "LLM" OR "LLMs" OR "foundation model*" OR "generative pre-trained transformer" OR "instruction-tuned" OR "ChatGPT" OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "LLaMA" OR "Mixtral" OR "Qwen" OR "DeepSeek" OR "Vicuna" OR "Pythia"
```

Concept block, Tier A (direction-locked):

```
"mechanistic interpretab*" OR "activation patching" OR "attribution patching" OR "causal tracing" OR "path patching" OR "sparse autoencoder*" OR "logit lens" OR "tuned lens" OR "circuit discovery" OR "induction head*" OR "polysemantic*" OR "residual stream" OR "steering vector*" OR "representation engineering" OR "probing classifier*" OR "linear prob*" OR "knowledge editing" OR "knowledge localization" OR "token attribution*" OR "input attribution*" OR "integrated gradient*" OR "saliency map*" OR "feature attribution*" OR "self-explanation*" OR "self-rationaliz*" OR "extractive rationale*"
```

Concept block, Tier B (generic; restricted to title or author-keyword fields):

```
"explainab*" OR "interpretab*" OR "XAI" OR "natural language explanation*" OR "counterfactual explanation*" OR "concept-based explanation*" OR "free-text rationale*" OR "shapley" OR "influence function*"
```

Bound-faithfulness block (corrective pass; `faithful*` qualifies only in combination with an explanation-context term):

```
"chain-of-thought faithfulness" OR "faithful chain-of-thought" OR "unfaithful" OR "faithfulness of explanation*" OR "faithful explanation*" OR "faithful rationale*" OR "faithfulness of reasoning" OR "post-hoc rationaliz*"
```

D.2 Scopus

Interface: scopus.com, Advanced Search. Dates run: 22 and 28 August 2026.

Tier A (adopted):

```
TITLE-ABS-KEY ( ( [P block] ) AND ( [Tier A concept block] ) ) AND PUBYEAR > 2018 AND PUBYEAR < 2027
```

Tier B, first version (superseded, n = 9,623):

```
TITLE-ABS-KEY ( ( [P block] ) AND ( [Tier B concept block] ) ) AND PUBYEAR > 2018 AND PUBYEAR < 2027
```

Superseded because generic vocabulary matched abstracts in which explainability appears only as motivation, yielding negligible precision.

Tier B, adopted version (n = 2,266):

```
TITLE-ABS-KEY ( [P block] )
AND TITLE ( "explainab*" OR "interpretab*" OR "XAI" OR "natural language explanation*" OR
    "counterfactual explanation*" OR "concept-based explanation*" OR
    "free-text rationale*" OR "shapley" OR "influence function*" )
AND PUBYEAR > 2018 AND PUBYEAR < 2027
```

Faithfulness top-up (n = 132):

```
TITLE-ABS-KEY ( ( "large language model*" OR "LLM" OR "LLMs" OR "ChatGPT" OR "GPT-4" )
    AND ( [bound-faithfulness block] ) ) AND PUBYEAR > 2018
```

EC1 volume query (planned to quantify the excluded adjacent literature; not screened): `TITLE-ABS-KEY (("large language model*" OR "LLM" OR "LLMs" OR "ChatGPT" OR "GPT-4") AND ("LLM-generated explanation*" OR "LLM-based explanation*" OR "conversational XAI" OR "explanation dialogue" OR "LLM-as-a-judge")) AND PUBYEAR > 2018 AND LANGUAGE(english) AND (SUBJAREA(COMP) OR SUBJAREA(ENGI))`

The result count of this query was not recorded, so the size of the excluded LLM-as-explainer literature is not reported.

Language, subject-area and document-type restrictions were applied through the interface refine panel in an earlier run and were not carried into the executed strings reported here. Records outside those criteria were removed at screening under codes X-LANG and X-TYPE.

D.3 ACM Digital Library

Interface: dl.acm.org, Advanced Search, ACM Full-Text Collection (858,371 records). Date run: 20 August 2026 according to the search record; the exported BibTeX files are time-stamped 27 August 2026 (Tier A/B) and 28 August 2026 (top-up). The ACM Guide to Computing Literature was not searched, to avoid duplicating other publishers' metadata without abstracts.

Tier A (n = 24 displayed; 25 exported): `Title:("large language model" OR "large language models" OR "LLM" OR "LLMs" OR "foundation model" OR "ChatGPT" OR "GPT-4" OR "LLaMA") AND Abstract:("mechanistic interpretability" OR "activation patching" OR "causal tracing" OR "sparse autoencoder" OR "logit lens" OR "residual stream" OR "probing classifier" OR "steering vector" OR "knowledge editing" OR "token attribution" OR "extractive rationale" OR "self-explanation")`

Tier B (n = 235 displayed; 238 exported): `Title:("large language model" OR "large language models" OR "LLM" OR "LLMs" OR "foundation model" OR "ChatGPT" OR "GPT-4" OR "LLaMA") AND Title:(explainability OR interpretability OR explainable OR interpretable OR "XAI" OR "natural language explanation" OR "counterfactual explanation" OR "influence functions" OR "Shapley")`

Faithfulness top-up (n = 6): `Title:("large language model" OR "large language models" OR "LLM" OR "LLMs" OR "ChatGPT" OR "GPT-4") AND Abstract:("chain-of-thought faithfulness" OR "faithful chain-of-thought" OR "unfaithful" OR "faithful explanation" OR "faithful rationale" OR "post-hoc rationalization")`

Publication-date filter 2019–2026 was requested. The interface date slider is bounded by the result set and did not extend below 2022 (Tier B) or 2024 (Tier A); no records were excluded by this bound.

D.4 ScienceDirect

Interface: sciencedirect.com. Date run: 27 August 2026. The interface permits a maximum of eight Boolean connectors per field and does not support truncation, so each tier was split across separate runs. Two field parameters are relevant: `qs` searches full text; `tak` searches title, abstract and keywords; `title` searches the title only.

Superseded: SD-1 executed under `qs` returned 103 records at negligible precision; SD-3 executed under `tak` returned more than 10,000. Both were re-run under a restricted field.

SD-1, Tier A, field `tak` (n = 14): `("large language model" OR "large language models" OR ChatGPT) AND ("mechanistic interpretability" OR "sparse autoencoder" OR "activation patching")`

SD-2, Tier A, field `tak` (n = 8): `("large language model" OR "large language models") AND ("probing classifier" OR "steering vector" OR "token attribution" OR "saliency map")`

SD-3, Tier B, field `title` (n = 119):
```
("large language model" OR "large language models" OR ChatGPT)
AND (explainability OR interpretability OR explainable OR interpretable)
```
SD-4, Tier B, field `title` (n = 0):
```
("large language model" OR "foundation model")
AND ("counterfactual explanation" OR "influence functions" OR "natural language explanation")
```
Filters: Years 2019–2026; article types Research articles and Review articles. ScienceDirect total: 141 records.

D.5 SpringerLink

Interface: link.springer.com, Advanced Search. Date run: 24 August 2026. The interface provides no abstract-restricted field, and its Keywords field matches full text rather than metadata. Field-prefixed Boolean syntax (`title:`) was not honoured by the current interface and returned no results; the form fields were used instead. Tier A was therefore **not executable at this source** and only Tier B was run.

Three passes were executed, each with a single phrase in the *Title* field and the same string in the *Keywords* field:

Title field: `large language model` (n = 495) • `LLM` (n = 889) • `ChatGPT` (n = 341). Union after DOI deduplication: 1,694.

Keywords field (all three passes):
```
interpretability OR explainability OR interpretable OR explainable OR
"mechanistic interpretability" OR "sparse autoencoder" OR "probing classifier" OR
"steering vector" OR "token attribution"
```
Filters: Discipline Computer Science; content types Article, Conference Paper, Chapter; English; 2019–2026; preview-only content excluded.

Because the Keywords field matched full text, the retrieved set had a precision below 6%. The Tier B rule (i.e. a population term *and* a generic explainability term in the title) was therefore applied locally after export, reducing 1,694 unique records to 100. Springer-published records are additionally captured at metadata precision via Scopus.

D.6 arXiv

Interface: arXiv Atom API (`export.arxiv.org/api/query`). Date run: 20 August 2026. Categories `cs.CL`, `cs.AI`, `cs.LG`, `cs.CY`, `stat.ML`; `submittedDate:[201901010000 TO 202612312359]`. Boolean operators uppercase; no wildcard support, so variants were enumerated. For query-length and pagination reasons the concept block was split into three sub-queries of ten terms, each executed once per annual `submittedDate` window (24 queries), with the template `(cat:cs.CL OR cat:cs.AI OR cat:cs.LG OR cat:cs.CY OR cat:stat.ML) AND (abs:P) AND (abs:`C_k`) AND submittedDate:[YYYY01010000 TO YYYY12312359]`. P and C are the lists in `arxiv_harvest.py`:

```
P: "large language model" OR "large language models" OR "foundation model" OR "foundation
models" OR "instruction-tuned" OR "generative pre-trained transformer" OR "ChatGPT" OR "GPT-3" OR
"GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "LLaMA" OR "Mixtral" OR "Qwen" OR "DeepSeek" OR "Vicuna" OR
"Pythia" OR "LLM" OR "LLMs"
C: "explainability" OR "explainable" OR "interpretability" OR "interpretable" OR "XAI" OR
"self-explanation" OR "natural language explanation" OR "saliency" OR "feature attribution"
OR "token attribution" OR "integrated gradients" OR "attention weights" OR "faithfulness" OR
"plausibility" OR "extractive rationale" OR "mechanistic interpretability" OR "activation patching"
OR "causal tracing" OR "sparse autoencoder" OR "dictionary learning" OR "probing classifier" OR
"logit lens" OR "influence functions" OR "counterfactual explanation" OR "concept-based explanation"
```

OR "Shapley" OR "steering vector" OR "representation engineering" OR "circuit analysis" OR "superposition"

The API does not enforce quoted phrase matching in the `abs:` field. The retrieved set of 13,685 records was therefore filtered locally to enforce the intended Boolean, requiring at least one population phrase and at least one concept phrase in title or abstract, yielding 7,888 records; the two-tier field restriction applied at all other sources then yielded 2,228 (Tier A 1,204; Tier B 1,024). The faithfulness rescue pass (28 August 2026) added 208. All figures are reported.

The rescue pass re-filtered the 20 August harvest, whose most recent record was submitted on 19 August, so submissions between 19 and 28 August 2026 were not covered by the original run. A supplementary search (`arxiv_gap_harvest.py`, run 3 October 2026) re-executed the three adopted sub-queries with `submittedDate:[202608190000 TO 202608282359]`. It retrieved 214 records (159, 76 and 3 per sub-query), of which 17 were already in the original harvest; the same local Boolean enforcement, tier classification and faithfulness rescue reduced the 197 new records to 81, 31 (Tier A 23; Tier B 8) and 1, so 32 records were adopted. Deduplication against the 5,698 records with the rules of Section 4.5 found no duplicates. The retrieval rate (21.4 records per day) matches that of the original run for 1–18 August 2026 (18.3 per day). The executed strings, counts and run time are in `arxiv_gap_run.json`.

D.7 ACL Anthology

The Anthology provides no Boolean search interface suitable for systematic retrieval but publishes complete metadata as a single BibTeX file (`anthology+abstracts.bib`, 130,930 entries; 74,692 within the 2019–2026 window). The same term classes and field restrictions used at all other sources were applied locally, yielding Tier A n = 435 and Tier B n = 264. Faithfulness rescue added 96.

D.8 IEEE Xplore

Interface: ieeexplore.ieee.org, Command Search. Date run: 20 August 2026. Filters: 2019–2026; Conferences, Journals and Magazines, Early Access Articles. The initial metadata-wide strategies (674 and 3,980 hits) were superseded because IEEE auto-assigned index terms dominated the yield; a Title × Abstract Tier B variant (1,509) was also superseded.

Tier A, adopted (n = 46):
```
("All Metadata":"large language model" OR "All Metadata":"large language models" OR "All Metadata":"LLM" OR "All Metadata":"LLMs" OR "All Metadata":"foundation model" OR "All Metadata":"ChatGPT" OR "All Metadata":"GPT-4" OR "All Metadata":"LLaMA")
AND ("Abstract":"mechanistic interpretability" OR "Abstract":"activation patching" OR "Abstract":"causal tracing" OR "Abstract":"sparse autoencoder" OR "Abstract":"logit lens" OR "Abstract":"residual stream" OR "Abstract":"probing classifier" OR "Abstract":"steering vector" OR "Abstract":"knowledge editing" OR "Abstract":"token attribution" OR "Abstract":"extractive rationale" OR "Abstract":"self-explanation")
```

Tier B, adopted (n = 567 displayed; 566 exported):
```
("Document Title":"large language model" OR "Document Title":"large language models" OR "Document Title":"LLM" OR "Document Title":"LLMs" OR "Document Title":"foundation model" OR "Document Title":"ChatGPT" OR "Document Title":"GPT-4" OR "Document Title":"LLaMA")
AND ("Document Title":explainab* OR "Document Title":interpretab* OR "Document Title":"XAI" OR "Author Keywords":explainab* OR "Author Keywords":interpretab* OR "Author Keywords":"XAI" OR "Author
```

```
Keywords":"natural language explanation" OR "Author Keywords":"counterfactual explanation" OR
"Author Keywords":"influence function" OR "Author Keywords":"Shapley")
```

D.9 Consistency with the exported records

Because platform search histories were not archived, the reported Scopus strings were checked against the exported records (title, abstract, author and index keywords) with a substring test. Of the 1,042 Tier A records, 1,035 contain at least one term of the Tier A concept block, and 89 contain `feature attribution` as their only concept term, whereas no record contains `superposition`, `attention weight` or `faithfulness` as its only concept term. The pattern of single-term records agrees with the string reported above and not with an earlier protocol draft that listed those three terms and omitted `feature attribution*`. The remaining seven records contain no Tier A term in any exported field, even after hyphens and punctuation are normalised. Scopus lemmatises search terms and searches fields that the export omits, so their retrieval cannot be reconstructed from the export; they are retained and screened like every other record. All 2,266 Tier B records contain a Tier B concept term in the title, and all 132 top-up records contain a term of the bound-faithfulness block. Moreover, the IEEE export files contain 46 and 566 records, the ScienceDirect exports 8 (SD-2), 14 (SD-1), and 100 and 19 records (SD-3, which the interface exported in two files) (141), the SpringerLink exports 495, 889 and 341 records (1,694 unique), the arXiv files 13,685, 7,888, 1,204 + 1,024 and 208 records, and the ACL Anthology tier files 435 and 264 records, matching the yields below.

D.10 Yield summary

D.11 Seed-set recall validation

Search recall was assessed against a seed set of fifteen papers chosen before the searches with the intention of spanning the method families; the identified seeds cover RF, FR, MI and BM only, and no seed was recorded for AT, PL, SA or HF. The seed set was previously reported as thirteen retrieved and two IC4 misses; however, identification of each seed in the deduplicated records (Table D4) does not fully support that report. Nine seeds match a unique record. One (Stolfo et al.) matches three records by the same first author, all of which are included in the census, so its outcome does not depend on which is meant. One (Paulo et al.) matches four records and cannot be resolved from the recorded shorthand, so it is removed and the smaller denominator is reported. Two (Conmy et al., SAEBench) have no matching record, also after the supplementary arXiv search of 2026-10-03, and are counted as search misses, in addition to the two reported IC4 misses. The seed set also motivated the corrective faithfulness pass and is therefore not an independent validation set. The fate of each seed under the superseded automated pilot, in which five in-scope seeds were removed by automation, is given in the released pilot record (`superseded_pilot/generated/tab_seeds.tex`).

Table D3 Search yields by source, including superseded strategies.

Source	Strategy	Raw	Adopted	Status
Scopus	Tier A	1,042	1,042	adopted
Scopus	Tier B (TITLE-ABS-KEY)	9,623	–	superseded
Scopus	Tier B (TITLE)	2,266	2,266	adopted
Scopus	faithfulness top-up	132	132	adopted
arXiv	API, phrase matching unenforced	13,685	–	superseded
arXiv	local Boolean enforcement	7,888	–	superseded
arXiv	local tier classification	2,228	2,228	adopted
arXiv	faithfulness rescue	208	208	adopted
arXiv	supplementary search, 19–28 Aug. 2026	214	32	adopted
ACL Anthology	local filter on BibTeX dump	699	699	adopted
ACL Anthology	faithfulness rescue	96	96	adopted
IEEE Xplore	Tier A (All Metadata)	674	–	superseded
IEEE Xplore	Tier A (All Metadata × Abstract)	46	46	adopted
IEEE Xplore	Tier B (All Metadata)	3,980	–	superseded
IEEE Xplore	Tier B (Title × Abstract)	1,509	–	superseded
IEEE Xplore	Tier B (Title × Title + Author Keywords)	567	566	adopted
ACM DL	Tier A	25	25	adopted
ACM DL	Tier B	238	238	adopted
ACM DL	faithfulness top-up	6	6	adopted
ScienceDirect	SD-1 (`qs`, full text)	103	–	superseded
ScienceDirect	SD-3 (`tak`)	>10,000	–	superseded
ScienceDirect	SD-1–SD-4 (restricted fields)	141	141	adopted
SpringerLink	three title passes, keywords full text	1,694	–	superseded
SpringerLink	local Tier B title rule	100	100	adopted
Total identified			7,825	

Table D4: Seed set: identification in the deduplicated records and decision in the census. Generated from `seed_set.csv` and the census decision files by `census_analysis.py`.

#	Seed (as recorded)	Record	D1	Retrieved by search	Census decision
1	Turpin et al., chain-of-thought unfaithfulness	Turpin et al. (2023)	RF	Yes	included

#	Seed (as recorded)	Record	D1	Retrieved by search	Census decision
2	Lanham et al., measuring CoT faithfulness	Lanham et al. (2023)	RF	Yes	included
3	Madsen et al.	Madsen et al. (2024)	RF	Yes	included
4	Gemma Scope	Lieberum et al. (2024)	FR	Yes	included
5	Llama Scope	He et al. (2024)	FR	Yes	included
6	Conmy et al.	–	MI	No	not in records
7	SAEBench	–	BM	No	not in records
8	Gao et al.	Gao et al. (2025)	FR	Yes	included
9	Stolfo et al.	–	?	Yes (ambiguous)	included, whichever of R00020, R00364, R02049 is meant
10	Shu et al., sparse autoencoder survey	Shu et al. (2025)	FR	Yes	excluded at title/abstract (X-SEC)
11	Paulo et al.	–	FR	–	removed: the shorthand matches four records and the intended paper cannot be identified
12	Bentham et al.	Bentham et al. (2024)	RF	Yes	included
13	Radhakrishnan et al.	Radhakrishnan et al. (2023)	RF	Yes	included
14	Wang et al., indirect object identification circuit	–	MI	No (IC4)	not in records
15	Belrose et al., tuned lens	–	MI	No (IC4)	not in records

Appendix E Exclusion Codes and Counts

The exclusion counts of the superseded automated pilot (rule-based policy, strict metadata screen, direction audit and pilot full-text gate) are given in the released pilot record for the automation audit; none of them is used in the census.

Table E5 Exclusion counts of the census by stage and code. Title/abstract counts are AI-reviewer exclusions (instructions v3) that R2 did not send forward; full-text counts are AI-reviewer exclusions (instructions v1), validated on R2's blind sample of 100 reports (Section 5.2); reports decided by R2 carry R2's decision. Exclusion codes are defined in the codebook, Section 1.

Stage	Code	Records
Title/abstract	X-DIR (LLM explains another system)	1,563
	X-TOPIC (not an explanation of an LLM)	1,221
	X-SEC (secondary study)	273
	X-MODEL (no LLM of $\geq$1B)	128
	X-TYPE (ineligible report type)	122
	X-TRANS (transparency without explanation)	7
	X-DUP (duplicate missed at deduplication)	1
	Total	3,315
Not retrieved	no open copy or institutional access	202
	scanned copy without a text layer	1
	Total	203
Full text	X-TOPIC	578
	X-MODEL	161
	X-DIR	166
	X-TYPE	52
	X-SEC	3
	Total	960
	Unclear, decided by R2	–

Appendix F Quality Appraisal Rubric

Each item is scored 1 (fully met), 0.5 (partly met) or 0 (not met); Q10 is NA when no human study is reported and is removed from the denominator. The normalised score is the sum divided by the number of applicable items.

Item	1	0.5	0
Q1 Objective and explanation claim	explicit claim naming explanandum and construct	objective stated, claim implicit	neither
Q2 Method replicability	all steps specified	main steps, details missing	not reproducible from text
Q3 Models identified	exact checkpoints and sizes	family and size	family only or none
Q4 Data, prompts, tasks	released or fully described	described, not released	not described
Q5 Code	link resolves to paper code	promised or partial	none

Item	1	0.5	0
Q6 Hyperparameters, seeds, decoding, compute	all four	two or three	at most one
Q7 Baselines and controls	baseline and random/shuffled control	baseline only	neither
Q8 Uncertainty	variance or CI over $\geq$3 runs or bootstrap	some error estimate	none
Q9 Faithfulness vs plausibility	evaluated separately or construct justified	one evaluated, other discussed	conflated
Q10 Human-study reporting (NA if none)	sample, task, protocol, measures, agreement	partial	not reported
Q11 Example selection	explicit criteria or random	partial rationale	showcase only
Q12 Limitations	specific failure modes	generic	none

Appendix G Deviations from the Original Protocol

1. The review is re-framed from a systematic area review to a systematic mapping study.
2. The IC1 preprint quality gate is withdrawn, and the encoder-only exclusion is moved into IC4.
3. Amendment A1 (frozen 2026-10-03, before any census decision): the dual human screening, assessment and coding of the protocol are replaced by a disclosed AI reviewer (R1) validated against one human reviewer (R2). The calibration gate ($\kappa \geq 0.70$) is retained; it failed with instructions v1 and v2 and passed with v3, after which all records were re-screened with v3; a third round and its consequences were pre-committed before it was run.
4. Calibration rounds 2 and 3 were reduced from 100 to 50 records for reasons of reviewer time, before R2 saw any record of these rounds.
5. The blind 20% random title/abstract sample (1,126 records) was replaced, before any record was drawn, by a random sample of 250 AI exclusions used to estimate the false-omission rate and screening recall; the safety rule that required R2 to screen every AI exclusion under an affected code was withdrawn as infeasible. Conservative recall (0.78) fell below the 0.95 threshold and is reported as a major limitation without further tuning.
6. Full-text contingency sampling is not applied: every retrievable eligible report is coded.
7. The blind 20% full-text sample (463 reports) and the per-field verification of all included studies were replaced on 2026-10-07, before the sample was drawn, by a blind random sample of 100 reports plus the recall-check records; fields with $\kappa < 0.60$ are flagged as unreliable. Two intermediate versions of R2's workbook filled by scripts driven by another AI tool were discarded, one derived from the AI reviewer's own decisions and one from keyword rules applied to the PDFs, because

neither is a human assessment; both are released for transparency and neither is used in any analysis.

8. The codebook was amended on 2026-10-03, before coding: D3 records the minimum access the method requires with a supporting quote, NON is assigned whenever FTH, PLA and HU are absent, `random_control` is set if and only if E1 is I2, and H2 requires ten or more participants.
9. Retrievability-based same-stratum replacement (pilot wave) is prohibited in the census.
10. The ordinal D5 scale is replaced by the four-field evidence profile, ROB is restricted to explanation robustness, and D1 records secondary codes for prevalence.
11. Citation searching, planned in the original protocol, was not performed.

References

Bao Y, Zhang X, Du T, et al (2025) Scalable Multi-Stage Influence Function for Large Language Models via Eigenvalue-Corrected Kronecker-Factored Parameterization. In: IJCAI International Joint Conference on Artificial Intelligence, pp 8022–8030, https://doi.org/10.24963/ijcai.2025/892

Bentham O, Stringham N, Marasović A (2024) Chain-of-Thought Unfaithfulness as Disguised Accuracy. arXiv preprint arXiv:2402.14897, arXiv:2402.14897

Bhambri S, Biswas U, Kambhampati S (2025) Interpretable Traces, Unexpected Outcomes: Investigating the Disconnect in Trace-Based Knowledge Distillation. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), https://doi.org/10.18653/v1/2026.acl-long.1686

Cambria E, Malandri L, Mercorio F, et al (2024) XAI meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models. arXiv preprint arXiv:2407.15248, arXiv:2407.15248

Chanin D, Hunter A, Camburu OM (2024) Identifying Linear Relational Concepts in Large Language Models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024, pp 1524–1535, https://doi.org/10.18653/v1/2024.naacl-long.85

Chen L, Dejl A, Toni F (2025a) Identifying Query-Relevant Neurons in Large Language Models for Long-Form Texts. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 23595–23604, https://doi.org/10.1609/aaai.v39i22.34529

Chen M, Engels J, Tegmark M (2025b) Low-Rank Adapting Models for Sparse Autoencoders. In: Proceedings of Machine Learning Research

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46. https://doi.org/10.1177/001316446002000104

van Dort I, Heuss M (2026) How Do LLMs Cite?: A Mechanistic Interpretation of Attribution in RAG. In: Lecture Notes in Computer Science, pp 458–473, https://doi.org/10.1007/978-3-032-21324-2_35

Duszenko J, Kazienko P, Kocoń J (2026) Sycophantic Anchors: Localizing and Quantifying User Agreement in Reasoning Models. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), https://doi.org/10.18653/v1/2026.acl-srw.20

European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024

Gao L, Tour T, Tillman H, et al (2025) SCALING AND EVALUATING SPARSE AUTOENCODERS. In: 13th International Conference on Learning Representations, ICLR 2025

Gao Y, Meng Q, Zhou Y, et al (2026) Towards Intrinsic Interpretability of Large Language Models: A Survey of Design Principles and Architectures. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, https://doi.org/10.18653/v1/2026.acl-long.1605

Hase P, Potts C (2026) Counterfactual Simulation Training for Chain-of-Thought Faithfulness. In: COLM 2026

He Y, Jiang Y, Wu Y, et al (2026) IDEA: An Interpretable and Editable Decision-Making Framework for LLMs via Verbal-to-Numeric Calibration. In: Findings of the Association for Computational Linguistics: ACL 2026, https://doi.org/10.18653/v1/2026.findings-acl.2004

He Z, Shu W, Ge X, et al (2024) Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. arXiv preprint arXiv:2410.20526, arXiv:2410.20526

Herrera F (2025) Making Sense of the Unsensible: Reflection, Survey, and Challenges for XAI in Large Language Models Toward Human-Centered AI. arXiv preprint arXiv:2505.20305, arXiv:2505.20305

Jacovi A, Goldberg Y (2020) Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 4198–4205, https://doi.org/10.18653/v1/2020.acl-main.386

Kang J, Butler L, Agarwal A, et al (2025) SPEX: Scaling Feature Interaction Explanations for LLMs. In: Proceedings of Machine Learning Research

Kastrati M, Imran A, Hashmi E, et al (2025) Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with Explainable Artificial Intelligence. Engineering Applications of Artificial Intelligence 149:110136. https://doi.org/10.1016/j.engappai.2025.110136

Kitchenham B, Charters S (2007) Guidelines for performing Systematic Literature Reviews in Software Engineering. Tech. Rep. EBSE-2007-01, Keele University and Durham University

Lanham T, Chen A, Radhakrishnan A, et al (2023) Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv preprint arXiv:2307.13702, arXiv:2307.13702

Li R, Gao Y (2025) Anchored Answers: Unravelling Positional Bias in GPT-2's Multiple-Choice Questions. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp 2439–2465, https://doi.org/10.18653/v1/2025.findings-acl.124

Lieberum T, Rajamanoharan S, Conmy A, et al (2024) Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In: Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, https://doi.org/10.18653/v1/2024.blackboxnlp-1.19

Lodh S, Obaidat I, Rustam F, et al (2025) Lightweight Fine-Tuning of LLMS for Explainable Intrusion Detection in SDN. In: International Conference on Wireless and Mobile Computing, Networking and Communications, https://doi.org/10.1109/WiMob66857.2025.11257572

Luo H, Specia L (2024) From Understanding to Utilization: A Survey on Explainability for Large Language Models. arXiv preprint arXiv:2401.12874, arXiv:2401.12874

Madsen A, Chandar S, Reddy S (2024) Are self-explanations from Large Language Models faithful? In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp 295–337, https://doi.org/10.18653/v1/2024.findings-acl.19

Makridis G, Fatouros G, Kampouris S, et al (2026) Review record of a systematic mapping study of explainability and interpretability methods for large language models. https://doi.org/10.5281/zenodo.23404135, URL https://doi.org/10.5281/zenodo.23404135, dataset

OpenAI (2023) GPT-4 Technical Report. arXiv preprint arXiv:2303.08774, arXiv:2303.08774

Page MJ, McKenzie JE, Bossuyt PM, et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71

Palikhe A, Yu Z, Wang Z, et al (2025) Towards Transparent AI: A Survey on Explainable Large Language Models. arXiv preprint arXiv:2506.21812, arXiv:2506.21812

Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology 64:1–18. https://doi.org/10.1016/j.infsof.2015.03.007

Polyakov G, Hepting C, Eickhoff C, et al (2025) Interpretability Analysis of Arithmetic In-Context Learning in Large Language Models. In: EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp 1758–1777, https://doi.org/10.18653/v1/2025.emnlp-main.92

Radhakrishnan A, Nguyen K, Chen A, et al (2023) Question Decomposition Improves the Faithfulness of Model-Generated Reasoning. arXiv preprint arXiv:2307.11768, arXiv:2307.11768

Rethlefsen ML, Kirtley S, Waffenschmidt S, et al (2021) PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Systematic Reviews 10:39. https://doi.org/10.1186/s13643-020-01542-z

Rimsky N, Gabrieli N, Schulz J, et al (2024) Steering Llama 2 via Contrastive Activation Addition. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp 15504–15522, https://doi.org/10.18653/v1/2024.acl-long.828

Sahoo S, Jain V, Chadha A, et al (2026) Linear Probes Detect Task Format, Not Reasoning Mode in Language Model Hidden States. In: Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), https://doi.org/10.18653/v1/2026.trustnlp-main.12

Saji A, Husain J, Jayakumar T, et al (2025) RomanLens: The Role of Latent Romanization in Multilinguality in LLMs. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp 26410–26429, https://doi.org/10.18653/v1/2025.findings-acl.1354

Shu D, Wu X, Zhao H, et al (2025) A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. In: EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2025, pp 1690–1712, https://doi.org/10.18653/v1/2025.findings-emnlp.89

Sovrano F (2025) Can Global XAI Methods Reveal Injected Behaviours in LLMs? SHAP vs Rule Extraction vs RuleSHAP. arXiv preprint arXiv:2505.11189, https://doi.org/10.1145/3770855.3818093, arXiv:2505.11189

Swaroop A, Nallani A, Uboweja S, et al (2025) FRIT: Using Causal Importance to Improve Chain-of-Thought Faithfulness. arXiv preprint arXiv:2509.13334, arXiv:2509.13334

Tu Z, Chen C, Du Y (2025) RRInf: Efficient Influence Function Estimation via Ridge Regression for Large Language Models and Text-to-Image Diffusion Models. In: EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp 18494–18507, https://doi.org/10.18653/v1/2025.emnlp-main.933

Turpin M, Michael J, Perez E, et al (2023) Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In: Advances in Neural Information Processing Systems

Valois P, Souza L, Shimomoto E, et al (2025) Frame Representation Hypothesis: Multi-Token LLM Interpretability and Concept-Guided Text Generation. Transactions of the Association for Computational Linguistics 13:1436–1458. https://doi.org/10.1162/TACL.a.48

Wilson EB (1927) Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22(158):209–212. https://doi.org/10.1080/01621459.1927.10502953

Zhang J, Lin Q, Rajmohan S, et al (2025) From Reasoning to Answer: Empirical, Attention-Based and Mechanistic Insights into Distilled DeepSeek R1 Models. In: EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp 3985–4002, https://doi.org/10.18653/v1/2025.emnlp-main.198

Zhang T, Deng Y, Jian P, et al (2026) Where CoT Reasoning Commits: Entropy Traces Identify Interpretable Attention Heads. In: Findings of the Association for Computational Linguistics: ACL 2026, https://doi.org/10.18653/v1/2026.findings-acl.133

Zhao H, Yang F, Shen B, et al (2024) Towards Uncovering How Large Language Model Works: An Explainability Perspective. arXiv preprint arXiv:2402.10688, arXiv:2402.10688