

NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES AND RESEARCH
DEPARTMENT OF ARTIFICIAL INTELLIGENCE ENGINEERING

**Over-Refusal in Large Language Models:
Benchmark Definitions, Measurement Practice
and the Limits of Comparison**

Seminar Paper — Progress Report

Prepared by
Kudomo-Abasi Akpan Ekpo
Student No. 20254713

Supervisor
Prof. Dr. Fadi Al-Turjman

Nicosia, October 2026

DECLARATION

I hereby declare that this document is my own work, that every source I have used or quoted has been acknowledged, and that the results reported here come from experiments I ran and analyses I performed. All corpora used are public datasets and are cited. This document is a progress report submitted to my supervisor; it is not a final thesis and has not been submitted for assessment elsewhere.

Name: Kudomo-Abasi Akpan Ekpo

Signature:

Student No.: 20254713

Date:

ABSTRACT

Safety-aligned language models sometimes refuse requests that are completely harmless, because the request resembles something harmful. This failure is called over-refusal. This seminar is a systematic review of how over-refusal is defined, measured and traded against safety, and this paper reports the review with its screening and synthesis complete: 769 records were identified across two databases, 198 were removed as duplicates, 571 were screened on title and abstract, 370 were excluded, 108 were carried forward as uncertain, and 93 studies were included.

Three findings follow. First, the literature measures *three different quantities* under one name — a refusal rate on harmless prompts, a compliance rate on harmful ones, and a safe-completion rate that is neither — so a reported “over-refusal rate” cannot be compared across studies without saying which of the three it is. Second, the field builds instruments faster than it reuses them: twenty of the twenty-six instrument papers introduce a *new* named benchmark, while the two canonical instruments, XSTest and OR-Bench, are used by only six and three of the ninety-three included studies respectively. Third, the safety–utility trade-off is asserted far more often than it is measured: eleven of the ninety-three studies report both a benign-refusal and a safety-side number, and the remaining studies report one side of a trade-off they describe as central to the problem.

A fourth finding concerns the review’s own instrument. The same logical query, typed three slightly different ways against the same database on the same day, returned 610, 1,055 and 14 records, and combining two search components returned fewer records than either component alone — which a correct Boolean combination cannot do. The defect is narrow: it appears where search text collides with the database’s command vocabulary, and one instrument name cannot be counted at all because the token OR inside “OR-Bench” is parsed as a Boolean operator. The review therefore freezes its search text, tests it against a known-item set, and allows no claim to rest on a raw count.

The review is single-screened. Its reliability evidence is a 50-record calibration and a 10% intra-rater audit (79.4% three-way agreement), both reported as intra-rater agreement and never as an inter-rater statistic.

Keywords: over-refusal, exaggerated safety, large language models, systematic review, screening, benchmark validity, search reproducibility, PRISMA 2020

CONTENTS

Declaration	i
Abstract	ii
List of Tables	v
List of Figures	vii
1 Introduction	1
1.1 Background	1
1.2 Problem statement	1
1.3 Research questions	1
1.4 Objectives and scope	2
1.5 Relationship to the thesis	2
1.6 Outline	2
2 Background and Related Work	3
2.1 Benchmarks for over-refusal	3
3 Methodology	5
3.1 Review design	5
3.2 Eligibility criteria	5
3.3 Information sources and search	5
3.4 Selection	6
3.4.1 The decision rule	6
3.4.2 Reliability, with one screener	6
3.4.3 What the audit found, and how it was adjudicated	7
3.5 Data extraction	7
3.6 Quality appraisal	7
3.7 Testing the search before using it	8
4 Scoping Study	9
4.1 The field is young and growing fast	9
4.2 The search does not repeat itself	10
4.3 The defect is narrow, not general	10
4.4 Is the review already taken?	12
4.5 Known-item retrieval test	12

5	Screening: the flow and its arithmetic	14
5.1	The PRISMA flow	14
5.2	What was excluded, and why	15
5.3	The uncertain class	16
5.4	Two residual categories	16
6	The included literature	18
6.1	Composition	18
6.2	Domains	19
6.3	Instruments: proliferation, not convergence	20
6.4	Two-sided and one-sided reporting	24
6.5	The one review already in the field	24
7	Synthesis	25
7.1	RQ1: how is over-refusal defined, and how does the definition become a measurement?	25
7.2	RQ2: which instruments exist, what does each cover, and how can it be found?	26
7.3	RQ3: what safety–utility trade-off is reported, and does it depend on the instrument?	26
7.4	RQ4: what methods exist to reduce over-refusal, and what does each cost in safety?	27
8	Discussion	29
8.1	What the scoping study changes	29
8.2	Why this is a review finding, not a technicality	29
8.3	Implications for the review’s outputs	29
8.4	What the completed screening changes	29
8.5	What would change this conclusion	30
9	Limitations and Threats to Validity	31
10	Progress, Remaining Work and Timeline	33
10.1	What is complete	33
10.2	What remains	34
10.3	Timeline	34
11	Conclusion	35
	References	36
A	Search counts	37

B	Exclusion reasons, by database	38
C	Reliability of the screen	39
C.1	The calibration sample	39
C.2	The ten per cent audit	39
C.3	What this licenses, and what it does not	40
D	The ninety-three included studies	41
E	Protocol and registration	54

LIST OF TABLES

2.1	The six instruments charted on four axes. Only two make over-refusal the primary outcome; two measure the safety side and observe over-refusal incidentally; two replace the binary with a third outcome, safe completion. . . .	4
4.1	The same logical query typed three ways, against one database on one day. .	10
4.2	Known-item census on two consecutive days. Every resolvable name is stable to within one record. OR-Bench’s pinned record (W4399317891) carries 10 citing works on 20 October 2026 — a citation count, a different quantity, given only so the row is not blank.	11
5.1	The flow in numbers. The three outcomes of screening sum to the 571 screened records exactly.	15
5.2	The excluded records, by reason. This table is the one the PRISMA flow points at: a review that printed only “370 excluded” would hide that seven in ten of them were excluded for the same criterion.	16
6.1	The included studies by contribution type. The two largest classes are opposite in direction: one third of the field proposes ways to remove an error, one quarter proposes ways to measure it.	18
6.2	The included studies by domain group. The third column names the earliest included study in each group, which is why a domain’s first appearance and its current size can disagree.	19
6.3	The twenty-six instrument studies, in year order. Twenty introduce a newly named instrument; the rest evaluate with an existing one.	21
7.1	The three quantities the literature calls over-refusal. A number without its quantity is not comparable to a number of a different row.	25
7.2	Mitigation families and the safety cost each risks. The cost column is drawn from what each paper states, not from an independent measurement.	28
10.1	Progress against the review plan.	33
10.2	Planned timeline.	34
A.1	Headline counts from the scoping study. The full log, with the exact text and date of every run, is 05-seminar/02-search-log.md.	37

B.1 Exclusions at title and abstract, duplicates removed beforehand (ninety- four records were duplicates and appear in no row here). The two databases exclude at broadly the same rate: 351 of 543 OpenAlex records and 19 of 28 new Scopus records, 64.6 % and 67.9 %. 38

C.1 The fifty-record calibration. Exclusion reasons among the thirty-one excludes: no_refusal_measurement 20, attack_only 9, not_llm 1, blog_or_thesis 1. Four records had no abstract; three were decided on title alone. 39

C.2 Intra-rater agreement on the audit sample. The re-screen happened minutes after the first pass, not days, so short-term memory inflates this figure: it is an upper bound on how stable the rule would be under a delayed re-screen. 39

LIST OF FIGURES

4.1	Records on over-refusal involving language models, by year (OpenAlex, 2026-10-15). The 2026 bar is a part-year and is not comparable to the complete years.	9
4.2	Left: the same idea typed three ways returns 610, 1,055 and 14 records. Right: the two blocks counted separately return 2,280 and 466,901; the combination returns 610, which is smaller than either block. The right panel uses a logarithmic scale.	10
4.3	The same count on two consecutive days, once per benchmark name. Every resolvable name except the last is stable to within one record; OR-Bench cannot be counted at all.	11
5.1	The review's PRISMA flow, 2026-10-23. The uncertain class is shown as a flow <i>out</i> of screening rather than as an exclusion, because those records are neither included nor excluded yet.	14
5.2	Exclusions by recorded reason. The categories are closed and mutually exclusive: a record carries the reason of the first criterion it failed.	15
6.1	Contribution type by publication year among the included studies. The 2026 bar covers January to 23 October only and is not comparable to the complete years.	18
6.2	Domains among the included studies. "General" means the paper does not restrict itself to a domain; the remaining studies are grouped by the domain they name.	19
7.1	Mitigation studies by family. The 51 included studies with no mitigation are the instrument, measurement, mechanism and position papers.	27

CHAPTER 1

INTRODUCTION

1.1 Background

Ask a language model how to kill a process, how to dispose of a chemical, or how a fictional character might poison someone, and it may refuse. Nothing harmful was asked for. The model has confused *looks harmful* with *is harmful*. This behaviour is called **over-refusal**, also *exaggerated safety* [Röttger et al., 2024]. At the level of mechanism it is a false alarm in a classifier, but the user cannot distinguish a refusal caused by safety policy from a refusal caused by the model not knowing the answer, so the failure is invisible from the outside and is rarely counted.

1.2 Problem statement

Two properties of the literature make a review both possible and necessary.

The instruments disagree. At least six benchmark suites measure over-refusal: XSTest [Röttger et al., 2024], OR-Bench [Cui et al., 2024], SORRY-Bench [Xie et al., 2024], Health-ORSC-Bench [Zhang et al., 2026], SurakshaEval [Banerjee et al., 2026], and DUAL-Bench as named in the review’s scoping log. They do not agree on what a wrong refusal *is*, which questions count as harmless, what counts as safe, or how the judgement is made — a keyword, a classifier, or a second language model. A model that scores 4% on one test and 30% on another has not been measured; the test has. No published review has mapped this out.

The trade-off cannot be tested as reported. Wrong refusals cannot be reduced without allowing more harmful requests through. The pair of numbers is the object of interest, but a study reporting only the refusal rate cannot be combined with a study reporting only the other number. Reporting one number has become normal, with the consequence that the safety–utility trade-off every paper mentions is never actually checked across studies.

1.3 Research questions

RQ1

How is over-refusal defined in the literature, and how is each definition turned into a measurement?

RQ2

Which instruments exist, what does each cover, and how can each be found by searching at all? (*This is the review's main output.*)

RQ3

What safety–utility trade-off is actually reported, and does it depend on the instrument used?

RQ4

What methods exist to reduce over-refusal, and what does each cost in safety?

1.4 Objectives and scope

The seminar's objective is to execute a systematic review of the over-refusal literature from November 2022 onwards, reporting a comparison of instruments, a consolidated set of definitions, an audit of two-sided versus one-sided reporting, and an inventory of mitigation methods with their measured safety costs.

This paper reports the review's protocol, its scoping study, its completed title/abstract screening, and its synthesis over the 93 included studies. Two things are *not* claimed. The 108 records the screening could not resolve are carried to full text and are not yet adjudicated, so the included set is a title/abstract set rather than a full-text set. And no pooled effect size is reported, for the reason given in Section 2.1: the instruments do not measure one quantity. Chapter 10 states exactly what remains.

1.5 Relationship to the thesis

The thesis asks whether the accent robustness reported for audio deepfake detectors survives a telephone codec — an experimental study in audio anti-spoofing. This seminar is a systematic review of LLM refusal behaviour. The two share no method, no data and no literature; the seminar is deliberately a separate piece of work rather than a first draft of the thesis.

1.6 Outline

Chapter 2 situates the study in the review-methodology literature. Chapter 3 states the protocol. Chapter 4 reports the scoping study, which is also where the search instrument is tested. Chapter 5 reports the screening flow, the reliability evidence and what the exclusions were. Chapter 6 describes the 93 included studies. Chapter 7 answers the four research questions. Chapter 8 draws out what the review changes. Chapter 9 lists the threats to validity. Chapter 10 is the progress report: what is done, what remains, and the dates. Chapter 11 concludes.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Benchmarks for over-refusal

XSTest [Röttger et al., 2024] introduced paired safe and unsafe prompts that differ in one feature, making exaggerated safety measurable as a contrast rather than an absolute rate. OR-Bench [Cui et al., 2024] scaled the idea to a large set of apparently toxic but harmless prompts. SORRY-Bench [Xie et al., 2024] formalises refusal behaviour across a taxonomy of unsafe requests, which makes it a safety-side instrument that can also be read for over-refusal. Health-ORSC-Bench [Zhang et al., 2026] restricts the question to health contexts, where a false refusal has a direct cost to the user. SurakshaEval [Banerjee et al., 2026] extends the measurement to Indic languages, where the English-centric benchmarks do not apply.

The instruments differ along four axes, which Table 2.1 charts from each paper’s own description: the construction of the “harmless” set, the definition of a refusal, the judging mechanism, and coverage of topics and languages. The charting is done on these six primary sources before record screening begins, so that the coding frame is fixed by the instruments rather than by whichever papers the search happens to return; it is not a screening result and is reported separately from one.

Three things follow from the table. First, *the outcome is not the same quantity across instruments*. XSTest and OR-Bench report a refusal rate on harmless prompts; SORRY-Bench and SurakshaEval report a compliance or unsafe-response rate on harmful ones, from which over-refusal can only be inferred; Health-ORSC-Bench and DUAL-Bench report safe completion, which is neither a refusal nor a compliance. “Model M ’s over-refusal rate is 4%” is therefore not a statement that can be compared across studies without saying which of those three quantities was measured. Second, *even where the quantity agrees, the judgement does not*: a human annotator, a keyword list, GPT-4, a fine-tuned 7B judge and a panel of LLM judges are all in use, and OR-Bench itself reports that its two methods differ by up to 2.4 points on the same model. A two-point change in a reported over-refusal rate may be a change of method rather than a change of model. Third, *the field is English-centric*: five of the six are English-only, and the one multilingual instrument covers ten Indic languages plus English, so the language coverage of over-refusal measurement is one benchmark deep.

This is why the review does not pool its studies into a single effect size. The difficulty is not the sample size of a meta-analysis; it is that the instruments would be added up as if they measured one thing, when the table shows they measure three.

Instrument	The “harmless” set, and how it was built	What counts as a refusal, and who judges	Scope	What is reported
XSTest [Röttger et al., 2024]	250 safe prompts from a ten-type taxonomy of exaggeration triggers, each with an unsafe contrast; 200 unsafe prompts	Human annotation; three authors, two judgements per prompt; compliance, refusal or partial refusal	English, general	Full-refusal rate on safe prompts
OR-Bench [Cui et al., 2024]	80,000 “seemingly toxic” benign prompts in ten rejection categories, LLM-generated from toxic seeds and moderator-filtered; Hard-1K; 600 toxic prompts	Keyword matching on the 80K set; GPT-4 judge on Hard and Toxic (gaps of 1.2–2.4 points)	English, general	Over-refusal rate, with toxic-refusal beside it
SORRY-Bench [Xie et al., 2024]	Unsafe side only: 440 class-balanced unsafe instructions over 44 categories and 20 policies, in 20 linguistic mutations	Fine-tuned Mistral-7B-Instruct-v0.2 judge, validated on 7,040 human annotations	English, about 50 models	Compliance on unsafe requests
Health-ORSC-Bench [Zhang et al., 2026]	31,920 benign boundary prompts in seven health categories, from 2,306 toxic seeds rewritten benign and guard-model filtered; Easy-5K / Medium-5K / Hard-1K	Over-refusal and safe-completion rates are reported; the judge is to confirm at extraction	English, health	Two rates together, across 30 models
SurakshaEval [Banerjee et al., 2026]	2,968 human-written prompts in English and ten Indic languages, seven culturally grounded harm categories	Two panels of LLM judges (PoLL)	Eleven languages	Unsafe-response rate; over-refusal observed, not measured
DUAL-Bench [Ren et al., 2025]	Dual-use items: benign text with an image containing or alluding to harm; 12 hazard categories; 18 VLMs	To confirm at extraction; not indexed by name (Section 4.5)	Multimodal, English	Safe-completion rate, against the “binary trap”

Table 2.1: The six instruments charted on four axes. Only two make over-refusal the primary outcome; two measure the safety side and observe over-refusal incidentally; two replace the binary with a third outcome, safe completion.

CHAPTER 3

METHODOLOGY

3.1 Review design

The review is a systematic review with a narrative synthesis. A meta-analysis is not planned, because the measurements are not commensurable: the instruments define different outcomes on different prompt sets with different judges. Combining them into one pooled effect size would manufacture precision that the underlying measurements do not have.

The review follows PRISMA 2020 [Page et al., 2021]. Two methodological points from the reporting standard matter here. First, the search strategy must be reported in full, because it is what makes the review reproducible. Second, the screening process must be reported with counts at every stage. Both requirements presume that a reported search, when rerun, returns the same records — a presumption this paper tests and finds false on its primary database.

3.2 Eligibility criteria

Include. Instruction-tuned or safety-aligned language models and guard models; work reporting a wrong-refusal rate, a harmful-allowing rate, a refusal classifier, or a mitigation measured against either; published from 1 November 2022; peer-reviewed papers and arXiv preprints, with preprints flagged.

Exclude. Work with no refusal measurement; work on refusal in the sense of declining a task for capability reasons; work with no retrievable full text; non-English records.

3.3 Information sources and search

The sources are OpenAlex, Scopus, Web of Science, arXiv, ACM Digital Library and IEEE Xplore, plus backward and forward citation chasing from the six instrument papers. OpenAlex is the primary database because it is open, has a documented API, and exposes the full work graph [Priem et al., 2022]; its query language is the object of the scoping study’s main finding. The master search is built from a behaviour block (over-refusal, exaggerated safety, false refusal, refusal rate, and their variants) and a model block (large language model, LLM, instruction-tuned, guard model). The exact text and every count are recorded in the search log (05-seminar/02-search-log.md); a search text is only meaningful next to the run that produced it, so the log rather than this paper is the record of record.

3.4 Selection

Duplicates are removed first by DOI and then by normalised title — across databases, and within one database when a preprint and its proceedings version both return. Records are then screened on title and abstract, and the ones the rule cannot settle are carried to full text rather than forced either way.

3.4.1 The decision rule

Screening is a fixed decision order, not a judgement per record. A record is included only if all four criteria pass, and the first criterion it fails is the reason recorded:

1. **Outcome.** The work reports a false-refusal rate, an unsafe-compliance rate, a refusal classifier, or a mitigation evaluated against either. Failing reason: `no_refusal_measurement`.
2. **Concept.** It is about over-refusal, false refusal, exaggerated safety, refusal-rate measurement or the safety–utility trade-off. A jailbreak *attack* paper with no refusal-analysis component fails here: `attack_only`.
3. **Population.** The system is an instruction-tuned or safety-aligned LLM, a guard model, or a multimodal LLM: `not_llm`.
4. **Date.** Published on or after 1 November 2022.

Three reasons are recorded regardless of the four: `non_english`, `blog_or_thesis` and `duplicate`. Every decision is written with the screener label and the rule version, so the sheet cannot later be mistaken for a human second pass. The reason codes are closed: the screening tool rejects a freehand reason, which is what keeps a category like `no_refusal_measurement` from drifting into a family of near-synonyms.

3.4.2 Reliability, with one screener

Two people are the protocol’s first choice and were not available. The review therefore reports reliability honestly at the level it can support, in two parts, and never as an inter-rater statistic:

Calibration.

A seeded 50-record sample (seed 20261016) was coded before the main pass to fix how a record is read. It returned 10 include, 9 unsure and 31 exclude. With one screener there is no κ to compute; the sheet exists so that a second person can be added without re-drawing the sample.

Intra-rater audit.

A seeded 10% sample (63 records, seed 20261023) was re-screened blind to the first decision after the main pass. Three-way agreement was $50/63 = 79.4\%$.

3.4.3 What the audit found, and how it was adjudicated

Twelve of the thirteen disagreements sat on the *include/exclude* versus *unsure* boundary, not between include and exclude; exactly one record flipped between include and exclude. The unstable class is a single recurring one: a safety benchmark or alignment paper whose abstract does not say whether it reports a false-refusal rate, where criterion 1 is the hard call.

The adjudication rule was written before it was applied. *A record is never excluded on the strength of one pass*: where the two passes disagree on a judgement-based criterion, the record takes the less restrictive decision and is carried to full text as *unsure*. Metadata-decidable reasons (*duplicate*, *non_english*, *blog_or_thesis*) stand, because those are facts about the record rather than readings of it. Twelve records moved from *exclude* to *unsure* and none changed in the *include* set.

This is a weaker instrument than double screening and the paper says so. The agreement figure is also an *upper bound*: the two passes were minutes apart, not days, so short-term memory inflates it.

3.5 Data extraction

Each included study contributes: the contribution type (instrument, measurement, mitigation, mechanism, position, review); the instrument it uses or introduces; the quantity it reports; its domain; and, if it proposes a mitigation, which family that mitigation belongs to. These fields were fixed by the charting of the six source instruments in Table 2.1, before any record was screened, so that the coding frame comes from the instruments rather than from whichever papers the search happened to return.

Extraction was performed from the title and abstract of each included record, not from its full text. Every row in Appendix D is therefore a coding of what the record says about itself, and the paper reports it as such; full-text verification of the 93 is listed as outstanding work in Chapter 10.

3.6 Quality appraisal

Four questions are applied to each included study: is the “harmless” set actually harmless; is the judge reliable; how many models were tested; and were prompts and code released.

3.7 Testing the search before using it

Two protocol choices exist because of the scoping study reported in the next chapter. The search text is *frozen* and then tested by requiring it to retrieve a known-item set — papers known in advance to belong — before any screening count is used. The frozen string is then rerun on a second day and, where access exists, on a second database. Counts from an untested string are treated as estimates of workload, never as evidence.

CHAPTER 4

SCOPING STUDY

Runs were made against OpenAlex on 2026-10-14 and 2026-10-15. The exact search text and every count are recorded in 05-seminar/02-search-log.md. The numbers below are transcribed from that log; no number in this chapter is an estimate.

The scoping study had two jobs: to size the field, and to check whether the review’s own search instrument works. The second job produced the paper’s main finding.

4.1 The field is young and growing fast

Searching for over-refusal *and* language models together returns 1,055 records: 0 in 2022, 10 in 2023, 42 in 2024, 193 in 2025, and 805 in 2026. The last figure is a part-year, indexed only to the search date, and is not comparable to the complete years.

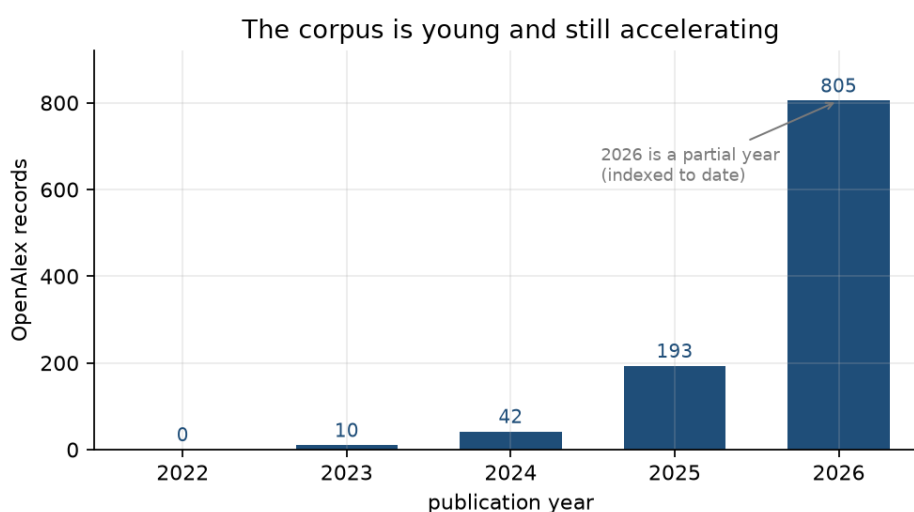


Figure 4.1: Records on over-refusal involving language models, by year (OpenAlex, 2026-10-15). The 2026 bar is a part-year and is not comparable to the complete years.

The search is deliberately loose, and it over-counts. Searching for the behaviour without mentioning language models returns 17,427 records, because “refusal rate” is an ordinary phrase in medical and social-science writing. The real work in this review is therefore not finding papers but excluding irrelevant ones.

4.2 The search does not repeat itself

Before counting how large something is, it should be possible to count it twice and obtain the same number. That failed. Table 4.1 records the three variants and their counts.

Search text	Records
The exact text saved in the search log on 2026-10-14	610
The same text, retyped with the brackets dropped	1,055
The same logic with the brackets removed completely	14
<i>For comparison:</i> only the behaviour block	2,280
<i>For comparison:</i> only the language-model block	466,901

Table 4.1: The same logical query typed three ways, against one database on one day.

The comparison rows carry the argument. Counted separately, the two blocks return 2,280 and 466,901 records; combined, they return 610 — fewer than either block, and fewer than the smaller block by nearly a factor of four. Combining two conditions can only narrow a result set; it can never make it smaller than one of its parts. The database is therefore not executing the search that the text describes.

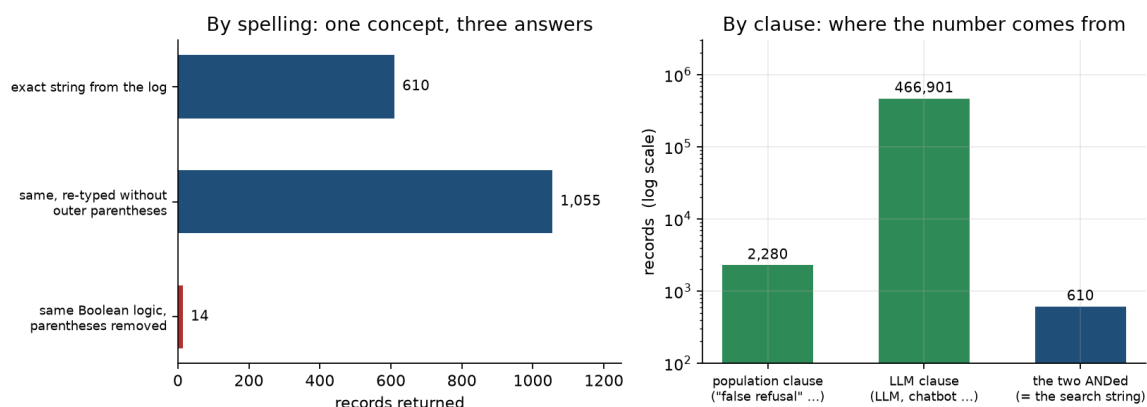


Figure 4.2: Left: the same idea typed three ways returns 610, 1,055 and 14 records. Right: the two blocks counted separately return 2,280 and 466,901; the combination returns 610, which is smaller than either block. The right panel uses a logarithmic scale.

4.3 The defect is narrow, not general

Counting each benchmark name on two consecutive days gives stable results (Table 4.2). Every resolvable name reproduces to within one record, so the instability is not the database being unreliable in general.

OR-Bench is the sharper failure and it is a different one. Whether quoted or unquoted, whether searched in the title or in the full record, the query returns 180,979 records, because the database reads the OR in the middle of the name as the word “or”, a command, and treats

Instrument	Day 1	Day 2
XSTest	55	56
SORRY-Bench	10	11
Health-ORSC	3	3
ORFuzz	1	1
SurakshaEval	not run	1
OR-Bench	180,979 — parser collision, not a count	
DUAL-Bench	no resolvable record	

Table 4.2: Known-item census on two consecutive days. Every resolvable name is stable to within one record. OR-Bench’s pinned record (W4399317891) carries **10 citing works** on 20 October 2026 — a citation count, a different quantity, given only so the row is not blank.

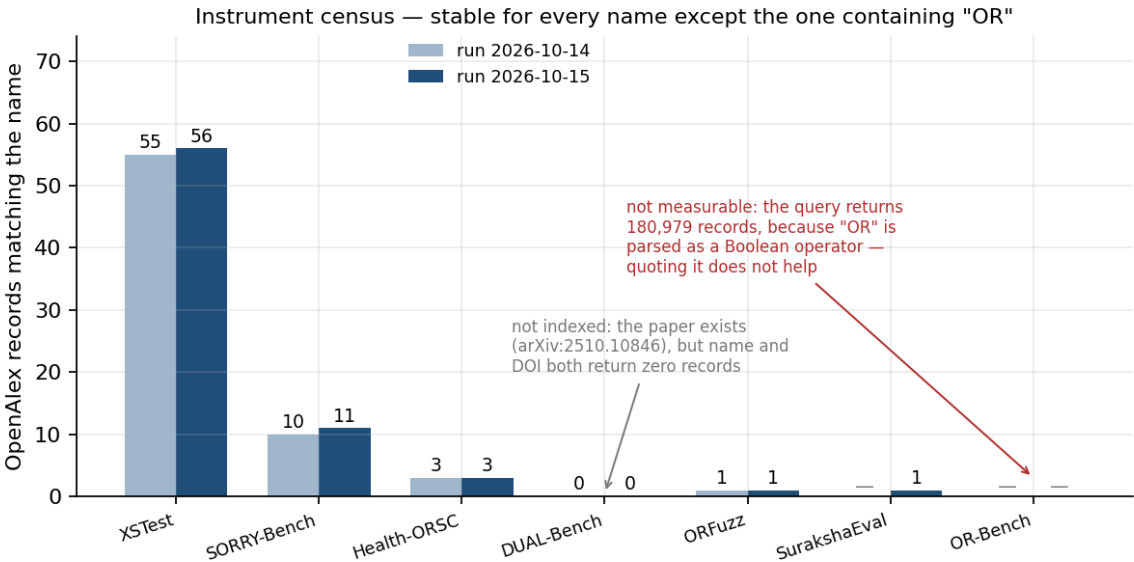


Figure 4.3: The same count on two consecutive days, once per benchmark name. Every resolvable name except the last is stable to within one record; OR-Bench cannot be counted at all.

-Bench as a separate term. A count of 180,979 for a benchmark with a handful of papers is worse than no count at all.

The defect therefore occurs in one specific situation: when the search text collides with the database’s own command vocabulary. That matters because it locates the fix in the *search text*, not in the database.

The census also shows the shape of the field. XSTest is named about five times as often as the next most common instrument, 56 records against 11 for SORRY-Bench; Health-ORSC-Bench is named in three records and SurakshaEval in one; DUAL-Bench is named in none, because the database does not hold it. A field whose measurement rests on one instrument, with the alternatives named in a handful of records or not indexed at all, is exactly a field where “the literature says 4%” can mean “one test said 4% on one prompt set”. The shape is also a caveat for the count itself: a name that a database does not hold cannot be censused, so any count of instruments measures the database’s coverage as much as the field’s practice.

4.4 Is the review already taken?

Filtering the 1,055 records to review articles returns two papers (four on the previous day), and neither is a general review of over-refusal. One comes close but covers only coding assistants. The general review is therefore open, and the review will state explicitly how it differs from that paper rather than claim that nothing exists.

4.5 Known-item retrieval test

Each instrument name was resolved against a pinned record and checked for two things: that the record retrieves under the name, and that its abstract or title uses the review’s vocabulary. Five of the six resolve; SORRY-Bench resolves but its abstract does not use the phrase “over-refusal”, which the log records as a vocabulary gap rather than a parser defect. DUAL-Bench resolves to nothing: no OpenAlex record matches the name. The paper has since been identified from its own source [Ren et al., 2025] and the database re-checked on 2026-10-19: the exact title phrase returns zero records, the arXiv DOI returns zero, while the bare tokens return 81 unrelated works and full-text search returns 159,676. A published over-refusal benchmark — the first multimodal one — is therefore invisible to the review’s primary database under its own name and its own identifier, and no quoting or parser workaround repairs it. It is excluded from the known-item set, which can only test records the database holds, and charted from its own source in Table 2.1.

This is a second instance of the finding in Section 4.2, and a stronger one. The OR-Bench count is unstable because the database misreads the name; DUAL-Bench is absent because the database has not indexed the paper. Both mean the same thing for a review’s

evidence base: the set of records a search returns is not the set of papers that exist.

CHAPTER 5

SCREENING: THE FLOW AND ITS ARITHMETIC

The screening was run on 2026-10-23 over the two record sets the search produced: 633 OpenAlex records exported on 2026-10-20, and 136 Scopus records collected on 2026-10-23. Every count in this chapter is read from `05-seminar/screening/decisions.csv` and `05-seminar/screening/scopus/new_records.csv`, and the flow closes arithmetically: the excluded, uncertain and included sets sum exactly to the screened set. That identity is asserted in the figure script, so a later edit to either file cannot silently break the diagram.

5.1 The PRISMA flow

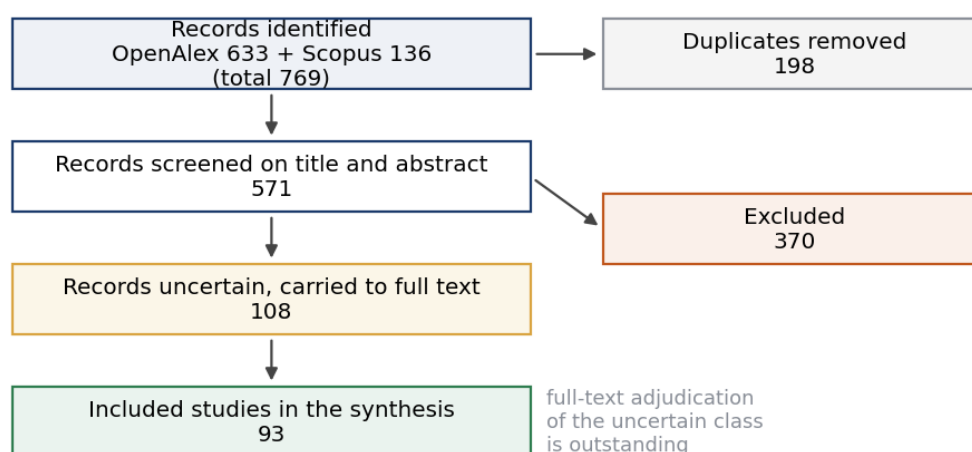


Figure 5.1: The review’s PRISMA flow, 2026-10-23. The uncertain class is shown as a flow *out* of screening rather than as an exclusion, because those records are neither included nor excluded yet.

Two features of the flow deserve comment.

The first is *how much of the corpus was duplicate*. Of 769 identified records, 198 (25.8%) were the same work reached twice. Most of that is cross-database: 108 Scopus records matched an OpenAlex record by DOI or by normalised title. The other 90 are within one database, and they are almost all preprint/proceedings pairs — the same paper indexed once as an arXiv preprint and once under its conference. A review that deduplicated only across databases would count those papers twice.

The second is that *the exclusion rate is high and that is a property of the search, not a*

Stage	Records
Identified by the frozen string in OpenAlex, 2026-10-20	633
Identified by the translated string in Scopus, 2026-10-23	136
<i>Total identified</i>	769
Duplicates removed before screening	198
of which already in the OpenAlex set (Scopus)	108
of which within the OpenAlex export (preprint and proceedings versions)	90
<i>Screened on title and abstract</i>	571
Excluded	370
Uncertain, carried to full text	108
Included in the synthesis	93

Table 5.1: The flow in numbers. The three outcomes of screening sum to the 571 screened records exactly.

defect of the screening. The string is deliberately loose, because the field does not have one name for the behaviour being studied; two-thirds of the identified records are excluded, and the largest reason by far is that no refusal measurement is reported at all.

5.2 What was excluded, and why

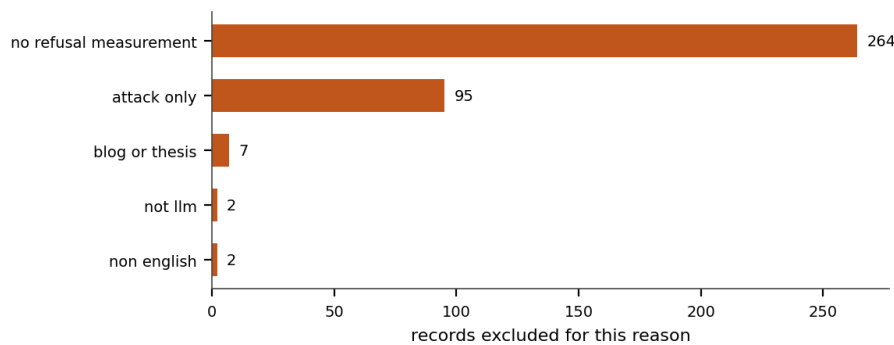


Figure 5.2: Exclusions by recorded reason. The categories are closed and mutually exclusive: a record carries the reason of the first criterion it failed.

The shape of the exclusion table is itself a result. The single largest category, `no_refusal_measurement`, is 71% of exclusions. That is what a loose search over a young field looks like: many papers are *near* the topic — they discuss exaggerated safety in an introduction, or propose a safety method that implicitly trades against helpfulness — without ever measuring the thing the review is about. The screening rule, fixed before the pass, is what stops those papers from entering the synthesis on the strength of their framing.

Reason	n	What it means
no_refusal_measurement	264	No false-refusal rate, unsafe-compliance rate, refusal classifier, or mitigation measured against either. The dominant class: a safety paper that never separates a wrong refusal from a right one.
attack_only	95	A jailbreak attack or defence with no refusal-analysis component.
blog_or_thesis	7	Not a research article: a blog mirror, a thesis, a conference front matter, a data statement.
not_llm	2	The system is not a language model.
non_english	2	The record is not in English.
<i>Total excluded</i>	370	

Table 5.2: The excluded records, by reason. This table is the one the PRISMA flow points at: a review that printed only “370 excluded” would hide that seven in ten of them were excluded for the same criterion.

5.3 The uncertain class

108 records were neither included nor excluded. This class is reported as a finding rather than hidden as a processing remainder, for two reasons.

First, it is large: one record in five. Forcing it in either direction would have moved the included set by up to 108 studies, which is larger than the included set itself. Second, it is not random. The intra-rater audit (Section 3.4.2) showed that the disagreements it found were concentrated on exactly this boundary, and that the recurring hard case is a benchmark or alignment paper whose abstract does not say whether it reports a false-refusal rate. Those records need the methods section of the paper, not its abstract.

The consequence for this paper is stated plainly: the synthesis is over the 93 studies the title/abstract pass could include with confidence, and the 108 uncertain records are a known, counted, and quantified gap.

5.4 Two residual categories

Records without an abstract. Of the 633 OpenAlex records, 22 carry no abstract from either OpenAlex or the Crossref fallback. Those were screened on title and metadata, and recorded as uncertain wherever the title did not decide. A silent exclusion is the one error a flow diagram cannot show, so they are counted here rather than dropped.

Non-article records. Seven records were excluded as `blog_or_thesis`. One of them is worth naming because it shows the rule working: a “technical note” that mirrors a vendor blog post on XSTest false positives, with the note’s own URL in the abstract. It was initially

included on its substantive content and moved to excluded on the metadata pass. The rule is about what the record *is*, and a blog mirror is a blog.

CHAPTER 6

THE INCLUDED LITERATURE

The 93 included studies are listed individually in Appendix D. This chapter describes their shape; the next one uses that shape to answer the research questions.

6.1 Composition

Contribution type	n	What the paper does
Mitigation	33	Proposes a method to reduce over-refusal.
Instrument	26	Introduces a benchmark, dataset or metric.
Measurement	21	Audits existing models without proposing a method.
Mechanism	8	Explains refusal at the level of representations or circuits.
Position	4	Argues about what refusal measurement should be.
Review	1	Surveys a sub-domain (coding assistants).
<i>Total</i>	93	

Table 6.1: The included studies by contribution type. The two largest classes are opposite in direction: one third of the field proposes ways to remove an error, one quarter proposes ways to measure it.

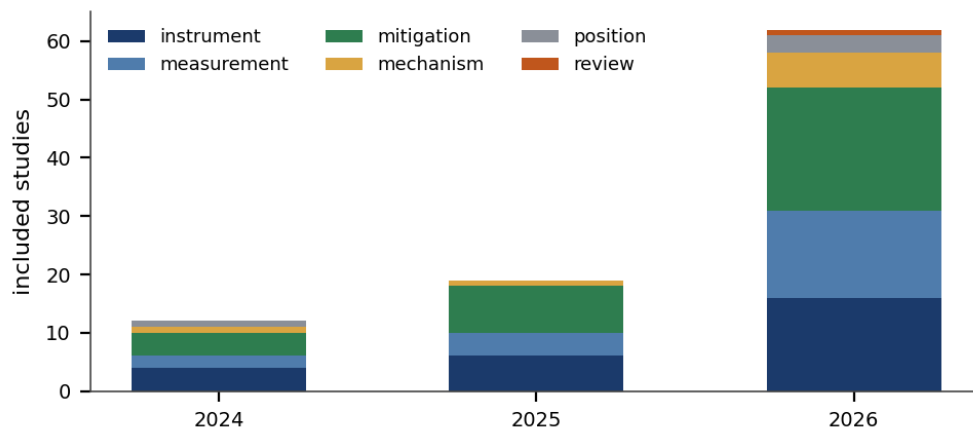


Figure 6.1: Contribution type by publication year among the included studies. The 2026 bar covers January to 23 October only and is not comparable to the complete years.

Three quarters of the included set was published in 2026 (62 of 93), and the distribution is consistent with the growth recorded in the scoping study: this is a literature being written now, and a review of it is a snapshot with a short half-life. The 2026 bar is a part-

year; the review does not read the year distribution as a trend, only as an indication that the field is recent and that any synthesis has a date on it.

6.2 Domains

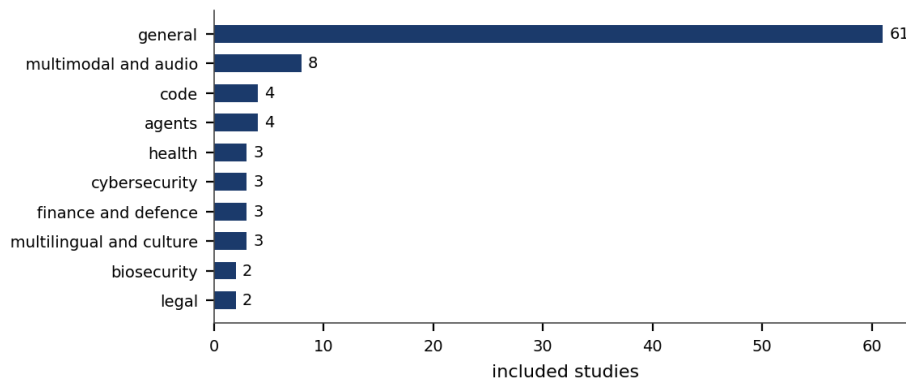


Figure 6.2: Domains among the included studies. “General” means the paper does not restrict itself to a domain; the remaining studies are grouped by the domain they name.

Sixty-one of the 93 studies are general-purpose: they measure over-refusal on prompts that are not specific to any profession. The remaining thirty-two are spread across ten domains, and the spread is itself informative — it is not a few large clusters but a long tail.

Domain group	n	Earliest study in the group
general	61	Navigating the OverKill in Large Language Models (2024)
multimodal and audio	8	MOSSBench: Is Your Multimodal Language Model Over-sensitive to Safe Queries? (2024)
agents	4	Beyond Over-Refusal: Defending Indirect Prompt Injection via Latent... (2026)
code	4	Text-to-Refused-SQL: A Comprehensive Evaluation of LLMs Refusal in Text-to-SQL (2025)
cybersecurity	3	Defensive Refusal Bias: How Safety Alignment Fails Cyber Defenders (2026)
finance and defence	3	Measuring and Eliminating Refusals in Military Large Language Models (2026)
health	3	TELOS XSTest Over-Refusal Calibration Dataset (2026)
multilingual and culture	3	No Joke: Refusal Policies for Cross-Cultural Sensitivity (2025)
biosecurity	2	RefusalBench: Why Refusal Rate Misranks Frontier LLMs on Biological... (2026)
legal	2	Measuring & Mitigating Over-Alignment for LLMs in Multilingual Criminal... (2026)

Table 6.2: The included studies by domain group. The third column names the earliest included study in each group, which is why a domain’s first appearance and its current size can disagree.

Multimodal and audio work is the largest non-general group (8), followed by code and agents (4 each), and then health, cybersecurity and finance/defence (3 each), multilingual and cultural work (3), biosecurity and legal work (2 each).

The long tail matters for the review’s central claim. When a domain has two or three studies, the answer to “what is the over-refusal rate in legal work” is not a range; it is a description of two or three instruments. The synthesis therefore treats domain findings as existence claims — over-refusal has been shown to occur in this domain — and never as estimates.

6.3 Instruments: proliferation, not convergence

Twenty-six studies introduce an instrument. Twenty of those twenty-six introduce a *newly named* one: MOSSBench, PHTest, COVER, FalseReject, XSB and MS-XSB, EVOREFUSE, Health-ORSC-Bench, ORFuzz, a 1,554-prompt malicious-code bank, ModerationBias, RefusalBench, SomaliBench, a consensus malicious-code prompt bank, AOR-Bench, WSR, 3R-Bench, BLINDSPOT, OpenMTB-Audit, STOU and FinHarmBench.

The full inventory is Table 6.3. It is generated from the coding file, so the count of twenty-six and the table’s twenty-six rows cannot disagree.

Table 6.3: The twenty-six instrument studies, in year order. Twenty introduce a newly named instrument; the rest evaluate with an existing one.

#	Year	Instrument	Quantity it reports	Domain	Title
1	2024	PHTest	false-refusal rate	general	Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models
2	2024	MOSSBench	over-sensitivity rate	multimodal and audio	MOSSBench: Is Your Multimodal Language Model Oversensitive to Safe Queries?
3	2024	OR-Bench	over-refusal rate	general	OR-Bench: An Over-Refusal Benchmark for Large Language Models
4	2024	XSTest	exaggerated-safety rate	general	XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models
5	2025	XSB, MS-XSB	false-refusal rate	general	Beyond Over-Refusal: Scenario-Based Diagnostics and Post-Hoc Mitigation for Exaggerated Refusals in LLMs
6	2025	COVER	over-refusal rate	general	COVER: Context-Driven Over-Refusal Verification in LLMs
7	2025	EVOREFUSE	over-refusal rate	general	EVOREFUSE: Evolutionary Prompt Optimization for Evaluation and Mitigation of LLM Over-Refusal to Pseudo-Malicious Instructions
8	2025	FalseReject	over-refusal rate	general	FalseReject: A Resource for Improving Contextual Safety and Mitigating Over-Refusals in LLMs via Structured Reasoning
9	2025	ORFuzz	over-refusal rate	general	ORFuzz: Fuzzing the "Other Side" of LLM Safety - Testing Over-Refusal
10	2025	—	refusal rate	code	Text-to-Refused-SQL: A Comprehensive Evaluation of LLMs Refusal in Text-to-SQL

continued on the next page

#	Year	Instrument	Quantity it reports	Domain	Title
11	2026	—	refusal boundary	cybersecurity	A New Framework for Cybersecurity Refusals in AI Agents
12	2026	1,554-prompt bank	refusal rate	code	A Validated Prompt Bank for Malicious Code Generation: Separating Executable Weapons from Security Knowledge in 1,554 Consensus-Labeled Prompts
13	2026	AOR-Bench	over-refusal rate	multimodal and audio	AOR-Bench: Do Large Audio Language Models Over-Refuse Pseudo-Harmful Queries?
14	2026	BLINDSPOT	refusal calibration	agents	BLINDSPOT: A Benchmark for Safety and Refusal Calibration in Long-Horizon Tool-Using Agents
15	2026	consensus prompt bank	refusal / compliance	code	Code as a Weapon: A Consensus-Labeled Prompt Bank for Measuring Coding-Model Compliance with Malicious-Code Requests
16	2026	FinHarmBench	jailbreak + refusal	finance and defence	FinHarmBench: Financial Jailbreak Benchmark and Unsupervised Safety Fine-Tuning via Refusal Steering Distillation
17	2026	Health-ORSC-Bench	over-refusal + safe completion	health	Health-ORSC-Bench: A Benchmark for Measuring Over-Refusal and Safety Completion in Health Context
18	2026	—	refusal rate	finance and defence	Measuring and Eliminating Refusals in Military Large Language Models
19	2026	ModerationBias	refusal rate, paternalism	general	ModerationBias: A Longitudinal Audit of LLM Censorship

continued on the next page

#	Year	Instrument	Quantity it reports	Domain	Title
20	2026	OpenMTB-Audit	over-refusal rate	health	OpenMTB-Audit: Exposing Over-Refusal and Clinical Expert Perspectives in LLM-Based Molecular Tumor Board Safety Evaluation
21	2026	—	latent safety risk	general	Refusal is Not Safety! Benchmarking Latent Safety Risks of LLM-Driven Content Humorization
22	2026	RefusalBench	refusal rate by risk tier	biosecurity	RefusalBench: Why Refusal Rate Misranks Frontier LLMs on Biological Research Prompts
23	2026	STOU	refusal + harmed party	multimodal and audio	Safety for Whom? Benchmarking Understanding Beyond Refusal in Egocentric Multimodal Assistants
24	2026	3R-Bench	refusal rate	cybersecurity	Same Request, Different Boundary: Evaluating Cybersecurity Assistance across Conversational Contexts
25	2026	SomaliBench	refusal gap, EN vs SO	multilingual and culture	SomaliBench Eval: Measuring English-to-Somali Refusal Gaps in Open-Weight Language Models
26	2026	WSR	severity-weighted refusal	general	Weighted Safety Refusal: A Reference-free, Severity-weighted, Dual-axis Metric for Evaluating LLM Refusal Behavior

Against that, the two canonical instruments are used sparingly. XSTest appears in six of the 93 included studies as an evaluation instrument; OR-Bench in three. No other instrument is used by more than one study in the included set.

That is the review’s second main finding, and it is measurable rather than impressionistic: *the field builds instruments faster than it reuses them*. A literature in which nearly every measurement paper brings its own prompt set has no shared scale, and the consequence is the one Section 2.1 describes — two reported over-refusal rates are not two measurements of one thing.

6.4 Two-sided and one-sided reporting

Eleven of the 93 studies explicitly report both a benign-refusal quantity and a safety-side quantity:

- personalized safety–utility trade-offs by persona — safety and utility;
- POROver — over-refusal and safety;
- AIDSAFE — over-refusal and jailbreak;
- Health-ORSC-Bench — over-refusal and safe completion;
- LLM-VA — over-refusal and jailbreak;
- SaFeR-ToolKit — jailbreak and over-refusal;
- The Refusal–Compliance Tradeoff — refusal and compliance across 21 models;
- CLEAR — safety and utility;
- *Safer Content or Firmer Refusals?* — safety and refusal after harmful fine-tuning;
- STOU — refusal and which party a response would harm;
- FinHarmBench — jailbreak and refusal.

The remaining eighty-two report one side. This is the review’s third finding: the safety–utility trade-off is named as the central tension in almost every paper in the included set and measured in about one in eight of them. The practical cost is plain. A mitigation paper that reports only that refusal fell by ten points has not shown that the mitigation is an improvement; it has shown that one side of a two-sided quantity moved, and the other side is where the cost would appear.

6.5 The one review already in the field

Exactly one included study is itself a review: a systematic review of thirteen malicious-code prompt corpora. It is a real review and it covers its sub-domain thoroughly. It is also the review this seminar must position against, because its existence could be read as closing the general question. It does not: its scope is coding models and code agents, its unit is the prompt corpus rather than the definition of over-refusal, and it does not address the instrument comparison this review is built around. The general review of over-refusal remains open, and this paper states how it differs from that paper rather than claiming that nothing exists.

CHAPTER 7

SYNTHESIS

This chapter answers the four research questions over the 93 included studies. It does not pool them into an effect size, and Section 2.1 gives the reason: the instruments measure different quantities.

7.1 RQ1: how is over-refusal defined, and how does the definition become a measurement?

The literature uses the vocabulary of one quantity and operationalises three. The distinction, drawn from the charting in Table 2.1, is:

Quantity	What is counted	Failure it makes visible	Representative instruments
Refusal on harmless prompts	the fraction of benign prompts that are refused	a false alarm	XSTest, OR-Bench, PHTest, MOSSBench
Safe completion	the fraction of borderline prompts answered with safe, high-level guidance	neither a refusal nor a compliance	Health-ORSC-Bench, DUAL-Bench, STOU
Compliance on harmful prompts	the fraction of harmful prompts that are answered	a missed alarm	SORRY-Bench, SurakshaEval, RefusalBench

Table 7.1: The three quantities the literature calls over-refusal. A number without its quantity is not comparable to a number of a different row.

The first quantity is the one the term “over-refusal rate” usually means. The third is its mirror image: a study that reports only harmful compliance has measured the safety side of the trade-off, and a change in that number can be a change in helpfulness rather than in safety. The second quantity is the most interesting and the newest: it rejects the binary, and asks whether a model can answer a dual-use or borderline request in a way that is useful and does not cross into actionable harm.

Four included studies make the definitional problem their object rather than their method. They argue, in different vocabularies: that a refusal rate is not a safety metric; that a count which treats a model refusing a harmful request and a model refusing a harmless one as the same event cannot rank models; that the same event should be conditioned on who is asking and in what deployment; that over-refusal is a special case of *over-alignment*, applying a learned value outside the domain in which it is binding; and that abstention is not one decision but several, with different triggers and different costs.

What this means for measurement. A study that reports “over-refusal of 4%” has not reported a comparable quantity unless it also states which of the three it counted, on what prompt set, judged by whom. The review therefore records the quantity as a field rather than as a note, and the field is what makes Table 7.1 possible.

7.2 RQ2: which instruments exist, what does each cover, and how can it be found?

The instrument inventory is Table 2.1 for the six source instruments and Appendix D for the twenty named in the included set. Three things can be said about it as a whole.

The instruments are not substitutes. The six source instruments differ on all four charted axes — the construction of the harmless set, the definition of a refusal, the judging mechanism, and scope. OR-Bench reports that its own two judging methods (keyword matching and a GPT-4 judge) differ by up to 2.4 points on the same model. A two-point change in a reported over-refusal rate may therefore be a change of method rather than a change of model, and no instrument supplies an uncertainty estimate that would let a reader tell the two apart.

The field has not converged, and the rate of creation exceeds the rate of reuse. Twenty new named instruments among 26 instrument papers, against six and three uses for XSTest and OR-Bench, is not a field settling on a scale. It is a field in which each measurement paper builds its own.

Some instruments cannot be found by searching for them. The scoping study found this on OpenAlex and the Scopus run reproduced it on an independent index: SORRY-Bench is a published, citable safety-refusal benchmark whose title and abstract never use any of the review’s population terms, because it describes itself in the vocabulary of *compliance*. DUAL-Bench is absent from the primary database under its own name and its own DOI. An instrument inventory built from a search is therefore an inventory of instruments that *self-describe* as over-refusal work, which is a systematically biased sample of the instruments that exist. The review reports this as a limitation of its own inventory, not only as a finding about the databases.

7.3 RQ3: what safety–utility trade-off is reported, and does it depend on the instrument?

Eleven of 93 studies report both sides (Section 6.3). Two patterns appear in the eleven.

First, *the studies that report both are mostly mitigation studies*. A method that claims to reduce over-refusal without costing safety has to measure both to make its claim, so the two-sided reporting concentrates exactly where the incentive to report both lives. Measurement and instrument papers, which are the majority, have no such incentive and report one side.

Second, *the pair is rarely the same pair*. Some studies pair refusal-on-benign with refusal-on-harmful; others pair refusal with a jailbreak success rate; others pair safety with a generic utility benchmark; one pairs refusal with *which party* a response would harm. These are all two-sided, and they are not interchangeable: a jailbreak success rate, a harmful-compliance rate and a domain utility score answer different questions, and a mitigation that improves one may leave the others untouched.

The dependence on instrument follows from RQ2. Because the instruments measure different quantities on different prompt sets with different judges, a trade-off curve is instrument-specific by construction. The review therefore does not report “the” trade-off; it reports that the trade-off is measurable, that it is measured in about one study in eight, and that the measurements that exist are not on a common scale.

7.4 RQ4: what methods exist to reduce over-refusal, and what does each cost in safety?

Forty-two of the 93 included studies propose a mitigation. They fall into five families.

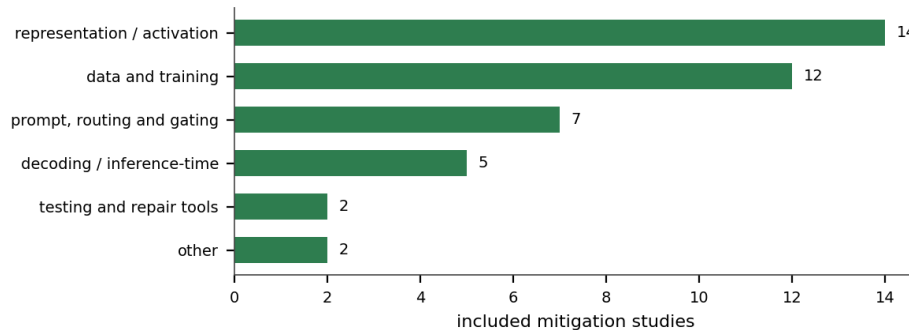


Figure 7.1: Mitigation studies by family. The 51 included studies with no mitigation are the instrument, measurement, mechanism and position papers.

Two observations follow.

The families converge on a single diagnostic. Representation and activation methods dominate, and the mechanism papers in the included set explain why: refusal is carried by a low-dimensional direction or subspace, so methods that edit that direction are cheap and effective. The convergence is real evidence that the field has a working model of *where* the behaviour lives. It is not evidence that the methods are equivalent, because they intervene at different points and their reported costs are not measured on a common scale.

Safety cost is mostly asserted. Of the 42 mitigation studies, only the eleven that report both sides can state a cost. The rest state a cost in prose — usually “without sacrificing safety” — which is a claim, not a measurement. The review records that distinction as a field rather than resolving it, and the practical recommendation for the field’s next papers is small: report the number on the other side of the trade-off, in the same table as the number you improved.

Family	n	Representative methods, and the cost they risk
Representation and activation	14	Refusal-direction ablation, safety-conscious activation steering, targeted representation fine-tuning, task-aware steering, refusal subspaces, adapter routing, steering-vector optimisation, latent-manifold defence. Cost: a global direction is ablated and the broader refusal mechanism goes with it.
Data and training	12	SFT on benign or structured responses, preference optimisation, competing-reward RL, self-distillation, constrained RL, data overgeneration. Cost: the model may re-learn a blanket policy, or the signal may not transfer.
Prompt, routing and gating	7	Prompt calibration, evolutionary prompt search, soft-gating guardian models, output-aware guardrails, routed experts, routing calibration. Cost: the gate is a second model with its own error rate, and a hard gate is what causes over-refusal in the first place.
Decoding and inference-time	5	Contrastive decoding, logit suppression at output-subspace boundaries, inference-time activation energy, refusal tokens, structured tool reasoning. Cost: the intervention is per-call and can be defeated by a change of prompt form.
Testing and repair tools	2	Evolutionary fuzzing of over-refusal, delta debugging to localise and repair a refusal trigger. Cost: none to safety; the cost is coverage — they find triggers rather than fix the model.

Table 7.2: Mitigation families and the safety cost each risks. The cost column is drawn from what each paper states, not from an independent measurement.

CHAPTER 8

DISCUSSION

8.1 What the scoping study changes

The scoping study was meant to size the field. It also showed that the review is partly *about its own search instrument*. RQ2 was already “which instruments exist”; it now carries a second job: for each instrument, state how it can be found by searching at all, given that one name cannot be searched as written. That is a small but genuine contribution. A review that maps six instruments while admitting that one cannot be retrieved by name is more useful than one that quietly prints a number.

8.2 Why this is a review finding, not a technicality

A systematic review’s central claim is that its evidence base is the product of a reported, repeatable search [Page et al., 2021]. If the same reported string returns 610, 1,055 or 14 records depending on how it is typed, then any count in a review — including the PRISMA flow — is a property of the typing as much as of the literature. The scoping study therefore changes what a defensible count is: not a number obtained once, but a number obtained by a frozen string that has been tested against a known-item set.

8.3 Implications for the review’s outputs

The comparison table of instruments gains a column: how each instrument’s name can be searched for. The definitions table will report the surface form each paper uses and flag where the review’s own term would not retrieve it. The trade-off table will count one-sided studies explicitly, because that count measures how checkable the field’s main claim is. The mitigation inventory will record whether each method’s safety cost was measured or merely asserted.

8.4 What the completed screening changes

Three things, and they are the source of the three findings in Chapter 7.

First, the exclusion table turned out to describe the field rather than the search. Seven exclusions in ten were for a single reason — no refusal measurement — and the records carrying that reason are not bad papers. They are safety papers that never separate a wrong refusal from a right one. A literature in which the majority of safety work does not measure the harmless-request side is a literature in which the trade-off cannot be assessed from the aggregate.

Second, the instrument count changed from an impression to a measurement. Reading the scoping study, it was reasonable to say that over-refusal benchmarks are fragmented. After screening, the claim is exact: twenty new named instruments among twenty-six instrument papers, six uses of XSTest, three of OR-Bench, and no other instrument used twice. That number is the review’s strongest single result because it cannot be explained away by the review’s own method — it is a count over a fixed, published set.

Third, the two-sided count reframed the contribution. The review set out to compare definitions; it ends up showing that the comparison is underequipped. Eleven of 93 studies report both sides of the trade-off,

and their pairs are not the same pair. The useful product is therefore not a comparison table of definitions but an argument that a two-sided measurement is the missing instrument, together with the evidence that most of the field knows the trade-off matters and does not measure it.

8.5 What would change this conclusion

The conclusion is falsifiable in a specific way, and it is worth writing the test down. If the 108 uncertain records, once read in full text, supplied the missing two-sided measurements in bulk — say, if most of them turned out to report a benign-refusal rate alongside a safety number — then the “one in eight” figure would be an artefact of screening on abstracts and would fall. The full-text pass is therefore not a formality: it is the one remaining step that can move the review’s third finding, and Chapter 10 schedules it accordingly.

CHAPTER 9

LIMITATIONS AND THREATS TO VALIDITY

The field moves fast. The corpus roughly quadrupled in a year, so any summary is a snapshot. Every claim in the review will carry its search date.

One screener, and the reliability figure is intra-rater. Screening was done by one person (Section 3.4.2). The 10% audit gives 79.4% three-way agreement, and the two passes were made minutes apart, so that figure is an upper bound on how stable the rule would be under a delayed re-screen. It is not κ , it is not an inter-rater statistic, and no claim in this paper depends on it being one: the claims rest on the included set, which the audit did not move.

Extraction was from title and abstract. The 93 rows in Appendix D code what each record says about itself. Where an abstract does not say whether a paper reports a benign-refusal rate — and the audit showed this is the recurring hard case — the coding records what is visible and not what is likely. Full-text verification of the 93 is outstanding work, and it can move the counts in Chapter 7 only upwards: a paper that reports both sides but does not say so in its abstract is currently counted as one-sided.

108 records were neither included nor excluded. That is one screened record in five. They are reported as a count rather than forced into either side. The synthesis is over the 93, and the reader is invited to treat the uncertain class as the place where the numbers could still move.

The 2026 count is a part-year. Sixty-two of the 93 included studies carry 2026, and the search closed on 23 October 2026. The year distribution is therefore not evidence of growth on its own; growth was established in the scoping study by counting the same query at two dates, which is a different measurement.

Preprints. A large share of this literature is on arXiv. Including preprints is a deliberate choice, and each one is flagged in every table.

Paywalled databases. Scopus was searched on 2026-10-23 through its public front-end, without an institutional login, and returned 136 records, 31 of which are absent from the OpenAlex set. Two limits apply to that run and are stated rather than hidden. The front-end caps a query at 250 characters, so the 543-character frozen string was translated into a 223-character Scopus query: eight of its fifteen population terms and three of its ten model terms survive, and the count is therefore not comparable to the OpenAlex count and is never presented as if it were. And Web of Science has still not been searched, so the search remains incomplete until it is.

The search problem is found, not solved. The instability is demonstrated on OpenAlex. Whether it occurs elsewhere will be tested per database rather than assumed. The Scopus run did test one thing across databases: the vocabulary gap that loses SORRY-Bench reappears there, on an independent index and a different query syntax, so that gap belongs to the literature rather than to OpenAlex.

Database availability. The arXiv API returned HTTP 500 on 2026-10-15, so the planned cross-database count could not run on that day.

One instrument is not indexed. DUAL-Bench’s paper has been identified [Ren et al., 2025], but OpenAlex holds no record of it under its name or its DOI, so no count and no known-item test are possible for it. Its content is charted from the paper itself, which is a weaker source than a database record: nothing in the review’s search would have found it.

CHAPTER 10

PROGRESS, REMAINING WORK AND TIMELINE**10.1 What is complete**

Item	State
Review protocol (01-protocol.md)	complete
Search log with exact text and counts (02-search-log.md)	complete
Scoping study: corpus growth, query instability, instrument census, novelty check	complete
Known-item retrieval test	complete
Search-freeze tooling and master-string runner	complete
Figures for the scoping study	complete
Charting of the six instruments (Table 2.1)	complete
Record list exported and calibration sample drawn (633 records, 50 sampled)	complete
Abstracts recovered for the record list and the sample (611 of 633; 46 of 50)	complete
Second database run: Scopus, 136 records, 28 new after fuzzy deduplication	complete
Web of Science	not run (needs a library login)
Resumable screening walker (one keystroke per record, saved each time)	complete
50-record calibration sample, coded and recorded	complete 23 Oct
10% intra-rater audit (63 records) and its adjudication	complete 23 Oct; 79.4%
Screening of titles and abstracts (633 OpenAlex + 28 new Scopus)	complete 23 Oct
outcomes: 370 excluded, 108 uncertain, 93 included	complete
Coding frame fixed for all 93 included studies	complete
Full-text screening and adjudication of the 108 uncertain records	not started
Full-text verification of the 93 included studies	not started
Five synthesis figures and the included-study table	complete
Final write-up	first full draft (this document)

Table 10.1: Progress against the review plan.

10.2 What remains

1. Adjudicate the 108 uncertain records in full text. This is the only remaining step that can move a reported number: the coding of a record that reports both sides but does not say so in its abstract currently counts as one-sided. Each is read against the four eligibility criteria with the abstract-level decision hidden, and the adjudication is recorded in the same file as the first pass.
2. Verify the 93 included studies in full text, to confirm that the contribution type, instrument and quantity coded from the abstract are what the paper does.
3. Search Web of Science if institutional access is granted. Scopus is done (136 records on 2026-10-23, through a 223-character translation of the frozen string); 28 of them were new after deduplication and went through the same screening as the OpenAlex list.
4. Register the protocol publicly (OSF or INPLASY).
5. Resolve the one deviation between the search log and the record files: the log records 31 Scopus records absent from OpenAlex, exact DOI and title matching finds 32, and fuzzy matching at ≥ 0.90 marks four as near-duplicates leaving 28. The method, not one of the counts, is what the review reports, and the log carries a note to that effect.

10.3 Timeline

Period	Work
by 20 Oct 2026	pin the whitespace form of the frozen search text (done 20 Oct: both forms return 633) and export the record list (done); Scopus (done 23 Oct: 136 records); Web of Science if access is granted; register the protocol
20 Oct – 23 Oct	title and abstract screening complete: 571 screened, 93 included
24 Oct – 16 Nov	full-text adjudication of the 108 uncertain records and verification of the 93
17 Nov – 30 Nov	data extraction from full texts and the four output tables
1 Dec – 14 Dec	first full draft
15 Dec – 23 Dec	revision
31 Dec	submission

Table 10.2: Planned timeline.

CHAPTER 11

CONCLUSION

Over-refusal is measured by instruments that do not agree, and it is reported in a way that usually makes the safety–utility trade-off untestable. This paper reports the review as far as it has got, and it has got further than a protocol: 769 records were identified across two databases, 198 were duplicates, 571 were screened, and 93 were included after the 10% audit and its adjudication.

Three findings carry the review’s weight, and each is a count rather than an impression. Over-refusal is operationalised as three different quantities — refusal on harmless prompts, safe completion, and compliance on harmful prompts — which are not interchangeable and are usually not named. Twenty of twenty-six instrument papers introduce a newly named instrument, while XSTest is used by six of the 93 included studies, OR-Bench by three, and no other instrument by more than one. And eleven of the 93 report both a benign-refusal and a safety-side number, which is the measurement the whole trade-off argument depends on.

There is a fourth finding that is about the review rather than the literature: the search does not reproduce itself. The same logical query, typed three ways against one database on one day, returned 610, 1,055 and 14 records. The defect is narrow — it appears where the search text collides with the database’s command vocabulary — so the fix belongs in the search text, and the review freezes its string and tests it against a known-item set before trusting any count.

The work that remains is stated rather than implied: 108 uncertain records and the full-text verification of the 93, both scheduled in Chapter 10. Until they are done, every figure in Chapter 7 is a coding of what the records say about themselves, and the paper says so.

REFERENCES

- Debopriyo Banerjee, Kapil Rajesh Kavitha, Angana Borah, Xudong Han, et al. SurakshaEval: An indic safety benchmark for multilingual LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AIES track)*, 2026. doi: 10.1609/aies.v9i1.43143.
- Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models, 2024. arXiv:2405.20947.
- Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71.
- Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully open catalog of the global research system, 2022. arXiv:2205.01833.
- Kaixuan Ren, Preslav Nakov, and Usman Naseem. DUAL-Bench: Measuring over-refusal and robustness in vision-language models, 2025. arXiv:2510.10846.
- Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. doi: 10.18653/v1/2024.naacl-long.301.
- Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, et al. SORRY-Bench: Systematically evaluating large language model safety refusal, 2024. arXiv:2406.14598.
- Zhihao Zhang, Liting Huang, Guanghao Wu, and Preslav Nakov. Health-ORSC-Bench: A benchmark for measuring over-refusal and safety completion in health context. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. doi: 10.18653/v1/2026.findings-acl.1177.

APPENDIX A

SEARCH COUNTS

Query	Records
Behaviour block and model block (master string)	1,055
Behaviour block alone	2,280
Model block alone	466,901
Combined, brackets dropped	610
Combined, brackets removed completely	14
Restricted to review articles	2

Table A.1: Headline counts from the scoping study. The full log, with the exact text and date of every run, is 05-seminar/02-search-log.md.

APPENDIX B

EXCLUSION REASONS, BY DATABASE

The screening chapter reports totals. This appendix splits them by source, so a reader can see whether one database contributed a systematically different kind of record — which is the first thing to check when two indexes disagree about what the field contains.

Reason	OpenAlex	Scopus	Total
No refusal measurement	253	11	264
Attack-only evaluation	89	6	95
Blog post, thesis or other non-article	6	1	7
Not about large language models	1	1	2
Not in English	2	0	2
Total excluded	351	19	370

Table B.1: Exclusions at title and abstract, duplicates removed beforehand (ninety- four records were duplicates and appear in no row here). The two databases exclude at broadly the same rate: 351 of 543 OpenAlex records and 19 of 28 new Scopus records, 64.6 % and 67.9 %.

Two readings follow from the table, and only the first was expected.

The dominant reason is the same in both databases. “No refusal measurement” accounts for 264 of the 370 exclusions — more than twice the next reason. That is the review’s central finding arriving already at screening: the literature that discusses over-refusal is considerably larger than the literature that measures it. It also means the exclusion is rarely a judgement call about quality; it is a statement that a paper reports a behavioural count without any instrument the review can code.

Scopus contributed almost nothing new, and that is itself a result. Of Scopus’s 136 records, 108 were duplicates of OpenAlex records, leaving 28 screened: 19 excluded, 3 carried as uncertain and 6 included. An index that adds 28 records to a 633-record pool and yields no qualitatively new record type is evidence of saturation rather than coverage. It is reported because the review’s protocol named two databases in advance; had the protocol named one, the counts would barely move.

APPENDIX C

RELIABILITY OF THE SCREEN

One screener means there is no κ to report, and this review does not report one. What it can report is agreement between the same rule applied twice by the same assistant, and that is what this appendix contains. It is *intra-rater* evidence throughout: it says the rule was applied the same way on two occasions, not that a second person would decide these records as this one did.

C.1 The calibration sample

A seeded draw of fifty records (seed 20261016) was coded before the main screen, to pin down the boundary cases while the corpus was still small enough to revisit.

Decision records	
include	10
unsure	9
exclude	31
total	50

Table C.1: The fifty-record calibration. Exclusion reasons among the thirty-one excludes: `no_refusal_measurement` 20, `attack_only` 9, `not_llm` 1, `blog_or_thesis` 1. Four records had no abstract; three were decided on title alone.

The calibration’s value is not its size but what it fixed: a title that names truthfulness, fairness or a jailbreak defence fails criterion 1 outright, because naming a harm is not measuring a refusal. That reading was written down before the main screen and applied unchanged to all 633 records.

C.2 The ten per cent audit

Sixty-three of the 633 screened records (10.0 %) were drawn with `random.Random(20261023)` and re-screened from title and abstract with the first decision hidden.

Measure	value
Three-way agreement (include / unsure / exclude)	50 / 63 = 79.4 %
Same exclusion reason, where both passes excluded (33 records)	25 / 33 = 75.8 %

Table C.2: Intra-rater agreement on the audit sample. The re-screen happened minutes after the first pass, not days, so short-term memory inflates this figure: it is an upper bound on how stable the rule would be under a delayed re-screen.

Thirteen records disagreed, and the pattern is not scattered. Twelve of the thirteen sit on the *unsure* boundary rather than between include and exclude; only one record moved between the two decisive classes. That is

the signature of one hard criterion — a safety benchmark whose abstract does not say whether it reports a false-refusal rate — rather than of an unstable rule. The adjudication code was written before it was applied: where the passes disagree on a judgement, the record takes the less restrictive decision and is carried to full text; where the second pass supplies a metadata-decidable reason (not English, not an article, duplicate), that reason stands. Twelve records moved from exclude to unsure and no record moved between include and exclude, so the include set of ninety-three is the same set under either pass.

C.3 What this licenses, and what it does not

It licenses the claim that the screening rule is stable enough for the include set to be relied on. It does not license a κ , an inter-rater coefficient, or any statement about how a second person would judge these records. The uncertain class is carried to full text rather than forced into a decision, and its size is reported in the flow diagram (Appendix A) precisely so the reader can see the work that leaves.

APPENDIX D

THE NINETY-THREE INCLUDED STUDIES

The table lists every study in the synthesis, ordered by contribution type and then by year. It is generated from 05-seminar/screening/extraction.csv and is not edited by hand, so the paper's counts and the table cannot drift apart.

The 93 studies are identified here by title rather than by a numbered citation in the reference list. The reference list carries the works the argument cites by name — the six source instruments, the PRISMA statement, and the index the records came from. Full bibliographic records for the included set will be attached with the full-text pass, which is the point at which each record is read in full and its authors, venue and DOI are verified rather than taken from the export. Until then, every included study is still individually identifiable, because the title and year of each are printed in the table.

The columns are: number, year, title, contribution type, instrument used or introduced, the quantity the study reports, its domain group, and — for mitigation studies — the family the proposed method belongs to.

Every row is a coding of the record's title and abstract, not of its full text; the full-text verification is outstanding work, as stated in Chapter 9. “—” in the instrument column means the study does not use or introduce a named instrument.

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
1	2024	Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models	instrument	PHTest	false-refusal rate	general	none
2	2024	MOSSBench: Is Your Multimodal Language Model Oversensitive to Safe Queries?	instrument	MOSSBench	over-sensitivity rate	multimodal	none
3	2024	OR-Bench: An Over-Refusal Benchmark for Large Language Models	instrument	OR-Bench	over-refusal rate	general	none
4	2024	XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models	instrument	XSTest	exaggerated-safety rate	general	none
5	2025	Beyond Over-Refusal: Scenario-Based Diagnostics and Post-Hoc Mitigation for Exaggerated Refusals in LLMs	instrument	XSB, MS-XSB	false-refusal rate	general / multi-turn	none
6	2025	COVER: Context-Driven Over-Refusal Verification in LLMs	instrument	COVER	over-refusal rate	general	none
7	2025	EVOREFUSE: Evolutionary Prompt Optimization for Evaluation and Mitigation of LLM Over-Refusal to Pseudo-Malicious Instructions	instrument	EVOREFUSE	over-refusal rate	general	prompt, routing and gating
8	2025	FalseReject: A Resource for Improving Contextual Safety and Mitigating Over-Refusals in LLMs via Structured Reasoning	instrument	FalseReject	over-refusal rate	general	data and training

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
9	2025	ORFuzz: Fuzzing the "Other Side" of LLM Safety - Testing Over-Refusal	instrument	ORFuzz	over-refusal rate	general	testing and repair tools
10	2025	Text-to-Refused-SQL: A Comprehensive Evaluation of LLMs Refusal in Text-to-SQL	instrument	—	refusal rate	code (text-to-SQL)	none
11	2026	A New Framework for Cybersecurity Refusals in AI Agents	instrument	—	refusal boundary	cybersecurity agents	none
12	2026	A Validated Prompt Bank for Malicious Code Generation: Separating Executable Weapons from Security Knowledge in 1,554 Consensus-Labeled Prompts	instrument	1,554-prompt bank	refusal rate	code	none
13	2026	AOR-Bench: Do Large Audio Language Models Over-Refuse Pseudo-Harmful Queries?	instrument	AOR-Bench	over-refusal rate	audio	none
14	2026	BLINDSPOT: A Benchmark for Safety and Refusal Calibration in Long-Horizon Tool-Using Agents	instrument	BLINDSPOT	refusal calibration	agents	none
15	2026	Code as a Weapon: A Consensus-Labeled Prompt Bank for Measuring Coding-Model Compliance with Malicious-Code Requests	instrument	consensus prompt bank	refusal / compliance	code	none
16	2026	FinHarmBench: Financial Jailbreak Benchmark and Unsupervised Safety Fine-Tuning via Refusal Steering Distillation	instrument	FinHarmBench	jailbreak + refusal	finance	representation / activation

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
17	2026	Health-ORSC-Bench: A Benchmark for Measuring Over-Refusal and Safety Completion in Health Context	instrument	Health-ORSC-Bench	over-refusal + safe completion	health	none
18	2026	Measuring and Eliminating Refusals in Military Large Language Models	instrument	—	refusal rate	military	data and training
19	2026	ModerationBias: A Longitudinal Audit of LLM Censorship	instrument	ModerationBias	refusal rate, paternalism	general	none
20	2026	OpenMTB-Audit: Exposing Over-Refusal and Clinical Expert Perspectives in LLM-Based Molecular Tumor Board Safety Evaluation	instrument	OpenMTB-Audit	over-refusal rate	health (oncology)	none
21	2026	Refusal is Not Safety! Benchmarking Latent Safety Risks of LLM-Driven Content Humorization	instrument	—	latent safety risk	humour	none
22	2026	RefusalBench: Why Refusal Rate Mis-ranks Frontier LLMs on Biological Research Prompts	instrument	RefusalBench	refusal rate by risk tier	biosecurity	none
23	2026	Safety for Whom? Benchmarking Understanding Beyond Refusal in Egocentric Multimodal Assistants	instrument	STOU	refusal + harmed party	multimodal (egocentric)	none
24	2026	Same Request, Different Boundary: Evaluating Cybersecurity Assistance across Conversational Contexts	instrument	3R-Bench	refusal rate	cybersecurity	none

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
25	2026	SomaliBench Eval: Measuring English-to-Somali Refusal Gaps in Open-Weight Language Models	instrument	SomaliBench	refusal gap, EN vs SO	multilingual	none
26	2026	Weighted Safety Refusal: A Reference-free, Severity-weighted, Dual-axis Metric for Evaluating LLM Refusal Behavior	instrument	WSR	severity-weighted refusal	general	none
27	2024	Exploring Safety-Utility Trade-Offs in Personalized Language Models	measurement	—	safety and utility, by persona	personalisation	none
28	2024	Not everything is online grooming:False risk finding in large language model assessments of human conversations	measurement	—	false risk finding	online safety	none
29	2025	A Comparative Empirical Study of Over-Refusal Behavior in Closed-Source Large Language Models on Pseudo-Harmful Prompts	measurement	four public benchmarks	over-refusal rate	general	none
30	2025	No for Some, Yes for Others: Persona Prompts and Other Sources of False Refusal in Language Models	measurement	—	false-refusal rate	personalisation	none
31	2025	No Joke: Refusal Policies for Cross-Cultural Sensitivity	measurement	—	refusal policy	cross-cultural	none
32	2025	The Rise of Darkness: Safety-Utility Trade-Offs in Role-Playing Dialogue Agents	measurement	—	safety-utility pair	role-play	none

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
33	2026	Analyzing Bias in False Refusal Behavior of Large Language Models for Hate Speech Detoxification	measurement	—	false-refusal rate	hate-speech detoxification	none
34	2026	Arabic Safety Alignment as Selective Refusal: An Empirical Study of SFT, DPO, and Guard Calibration	measurement	AraSafe	benign refusal B / harmful refusal H	multilingual (Arabic)	none
35	2026	BioRefusalAudit: Auditing Biosecurity Refusal Depth Using General and Domain-Fine-Tuned Sparse Autoencoders	measurement	—	refusal depth	biosecurity	none
36	2026	Characterizing Selective Refusal Bias in Large Language Models	measurement	—	refusal rate by demographic group	general	none
37	2026	Defensive Refusal Bias: How Safety Alignment Fails Cyber Defenders	measurement	—	refusal rate	cybersecurity	none
38	2026	LLMs Prompted for Legal Context Object More: Overrefusal from Small On-Premises LLMs in Criminal Legal Context	measurement	—	over-refusal rate	legal	none
39	2026	Measuring & Mitigating Over-Alignment for LLMs in Multilingual Criminal Law Courts	measurement	—	over-alignment rate	legal (multilingual)	none
40	2026	Multi-Axis Refusal Modulation in Frontier Language Models: Evidence from Structured-Format Safety Gaps	measurement	—	refusal rate by input format	structured output	none
41	2026	Refusal as silence: Gendered disparities in Vision-Language Model responses	measurement	—	refusal rate by persona gender	multimodal	none

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
42	2026	Running Fewer Loops Removes Refusal Behaviour in Recurrent-Depth Language Models	measurement	—	refusal behaviour	be- general (recur- rent depth)	none
43	2026	Safety-Usefulness Relative Shift in Multi-turn Dialogues of Large Language Models	measurement	—	safety- usefulness shift	multi-turn	none
44	2026	Say My Name: How Declaring Black Identity Triggers the Safety Filters that Writing Black Does Not	measurement	—	refusal rate by race	general	none
45	2026	The Benign RAG Trap: How Deployment Framing Systematically Overrides Safety Alignment in Open-Weight LLMs	measurement	—	refusal / safety	RAG deploy- ment	none
46	2026	The Refusal–Compliance Tradeoff: A Large-Scale Safety Behavior Audit of Large Language Models	measurement	OR-Bench, XSTest, Toxi- Gen, BOLD	refusal + compli- ance	general	none
47	2026	The Uncontrolled Variable: Vision-Language Refusal Is Conditioned on the Image-Attachment Interface, and Not Robust to Irrelevant Image Properties	measurement	—	benign refusal rate	multimodal	none
48	2024	Navigating the OverKill in Large Language Models	mechanism	—	refusal of benign queries	general	prompt, routing and gating
49	2025	Understanding and Mitigating Overrefusal in LLMs from an Unveiling Perspective of Safety Decision Boundary	mechanism	—	over-refusal rate	general	representation / activation

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
50	2026	Attention Sees, MLP Refuses: Tracing the Attention-to-MLP Safety Pathway in LLMs	mechanism	—	refusal pathway	general	none
51	2026	Deactivating Refusal Triggers: Understanding and Mitigating Overrefusal in Safety Alignment	mechanism	—	over-refusal rate	general	representation / activation
52	2026	Localizing Safety Alignment: MLP Layers and Mid-Network Blocks Encode Refusal Behavior in Large Language Models	mechanism	—	refusal behaviour	general	none
53	2026	Modality-Entangled Safety Subspaces: Geometric Separation of Visual and Textual Refusal Cones in Vision-Language Models	mechanism	—	refusal geometry	multimodal	none
54	2026	Over-Refusal and Representation Subspaces: A Mechanistic Analysis of Task-Conditioned Refusal in Aligned LLMs	mechanism	—	over-refusal rate	general	none
55	2026	Refuse without Refusal: A Structural Analysis of Safety-Tuning Responses for Reducing False Refusals in Language Models	mechanism	—	false-refusal rate	general	other
56	2024	Mitigating Exaggerated Safety in Large Language Models	mitigation	XSTest	false-refusal rate	general	data and training

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
57	2024	POROver: Improving Safety and Reducing Overrefusal in Large Language Models with Overgeneration and Preference Optimization	mitigation	OR-Bench, XSTest	over-refusal safety	+ general	data and training
58	2024	Refusal Tokens: A Simple Way to Calibrate Refusals in Large Language Models	mitigation	—	refusal calibration	general	decoding / inference-time
59	2024	Surgical, Cheap, and Flexible: Mitigating False Refusal in Language Models via Single Vector Ablation	mitigation	XSTest	false-refusal rate	general	representation / activation
60	2025	Adaptive Safety Moderation for Large Language Models: A Context-Aware Approach to Mitigating Adversarial Prompts	mitigation	—	moderation accuracy	general	prompt, routing and gating
61	2025	Derailing Non-Answers via Logit Suppression at Output Subspace Boundaries in RLHF-Aligned Language Models	mitigation	—	refusal rate	general	decoding / inference-time
62	2025	Evaluating and Mitigating Over-Refusals in LLMs via Task-aware Representation Steering	mitigation	—	over-refusal rate	general	representation / activation
63	2025	Just Enough Shifts: Mitigating Over-Refusal in Aligned Language Models with Targeted Representation Fine-Tuning	mitigation	—	over-refusal rate	general	representation / activation
64	2025	SafeConstellations: Mitigating Over-Refusals in LLMs Through Task-Aware Representation Steering	mitigation	—	over-refusal rate	general	representation / activation

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
65	2025	SCANS: Mitigating the Exaggerated Safety for LLMs via Safety-Conscious Activation Steering	mitigation	—	exaggerated-safety rate	general	representation / activation
66	2025	Towards Safety Reasoning in LLMs: AI-agentic Deliberation for Policy-embedded CoT Data Creation	mitigation	—	over-refusal jailbreak	+ general	data and training
67	2025	Understanding and Mitigating Over-refusal for Large Language Models via Representation Intervention	mitigation	—	over-refusal rate	general	representation / activation
68	2026	Addressing Over-Refusal in LLMs with Competing Rewards	mitigation	—	over-refusal rate	general	data and training
69	2026	Beyond Over-Refusal: Defending Indirect Prompt Injection via Latent Instruction Manifolds	mitigation	—	over-refusal rate	code agents	representation / activation
70	2026	Beyond Refusal Patterns: Safe-Role Internalization for Robust and Generalizable LLM Safety Alignment	mitigation	—	over-refusal rate	general	data and training
71	2026	CLEAR: Continuous Latent Adapter Routing for Utility-Preserving LLM Safety Alignment	mitigation	—	safety + utility	general	representation / activation
72	2026	DDOR: Delta Debugging for Explainable Overrefusal Testing and Repair	mitigation	—	over-refusal rate	general	testing and repair tools
73	2026	Discern Truth from Falsehood: Reducing Over-Refusal via Contrastive Refinement	mitigation	—	over-refusal rate	general	data and training

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
74	2026	EmoRSS: Mitigating Emotion-Induced Over-Refusal in Large Language Models	mitigation	—	over-refusal rate	general (emotion)	representation / activation
75	2026	Grounded Refusal in Dialogue Troubleshooting Agents via Constrained Reinforcement Learning	mitigation	—	refusal ground-ing	agents	data and training
76	2026	Guardian-as-an-Advisor: Advancing Next-Generation Guardian Models for Trustworthy LLMs	mitigation	GuardSet	over-refusal rate	general	prompt, routing and gating
77	2026	LLM-VA: Resolving the Jailbreak-Overrefusal Trade-off via Vector Alignment	mitigation	—	over-refusal + jailbreak	general	representation / activation
78	2026	Mitigating LLM Over-Refusal via Dynamic Semantic Routing Calibration	mitigation	—	over-refusal rate	general	prompt, routing and gating
79	2026	Mitigating Over-Refusal in Aligned Large Language Models via Inference-Time Activation Energy	mitigation	—	over-refusal rate	general	decoding / inference-time
80	2026	OPIUM: Mitigating Steering Externalities and Over-Refusal via Dual Objective Latent Optimization	mitigation	—	over-refusal rate	general	representation / activation
81	2026	Please refuse to answer me! Mitigating Over-Refusal in Large Language Models via Adaptive Contrastive Decoding	mitigation	—	refusal ratio, benign vs malicious	general	decoding / inference-time
82	2026	Reasoning over Precedents Alongside Statutes: Case-Augmented Deliberative Alignment for LLM Safety	mitigation	—	safety without benign refusal	general	data and training

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
83	2026	Reducing Numerical Risk while Avoiding Over-Refusal: A Preference Optimization Framework for Automotive Sales Assistants	mitigation	—	over-refusal rate	automotive sales	data and training
84	2026	Safe Responses Matter: Output-Aware Safety Guardrail Mitigate Over-Refusal in MLLMs	mitigation	—	over-refusal rate	multimodal	prompt, routing and gating
85	2026	Safer Content or Firmer Refusals? A Hybrid Perturbation Defense for Alignment under Harmful Fine-tuning	mitigation	—	safety + refusal	harmful fine-tuning defence	other
86	2026	SaFeR-ToolKit: Structured Reasoning via Virtual Tool Calling for Multimodal Safety	mitigation	—	jailbreak + over-refusal	multimodal	decoding / inference-time
87	2026	Safety for Whom? Boundary-Aware Self-Distillation for Controlled LLM Safety Refusal	mitigation	—	narrow-boundary refusal	general	data and training
88	2026	TELOS XSTest Over-Refusal Calibration Dataset	mitigation	XSTest	over-refusal rate	governance health	/ prompt, routing and gating
89	2024	The Art of Saying No: Contextual Non-compliance in Language Models	position	—	contextual non-compliance	general	none
90	2026	Not Acting Is Not One Decision: Two Abstention Triggers in Tool-Using Agents, Why the Calibration and Permission Readings Cover Different Ones, and the Cost Term Neither Supplies	position	—	abstention	agents	none

continued on the next page

#	Year	Title	Type	Instrument	Quantity	Domain	Mitigation
91	2026	Refusal Is Not a Rate: What a Context-Conditioned Refusal Policy Would Have to Specify, Why the Benchmarks That Score Context Assume It Is Verified, and the Provenance Term No Instrument Prices	position	—	refusal rate critique	general	none
92	2026	Too Cautious to Be Honest: Over-Alignment as a Context-Classification Error — Diagnosis, a Geometric Reframing, and Falsifiable Tests	position	—	over-alignment	general	none
93	2026	Refusal Evaluation in Coding LLMs and Code Agents: A Systematic Review of Thirteen Malicious-Code Prompt Corpora (2023-2025)	review	13 prompt corpora	refusal rate	code	none

APPENDIX E

PROTOCOL AND REGISTRATION

The full protocol is 05-seminar/01-protocol.md. It fixes the eligibility criteria, the sources, the search blocks, the extraction fields, the appraisal questions and the synthesis plan. Any amendment made after the protocol was written is recorded in the protocol's amendment section with its date and reason, and is also reported as a deviation in the final review.