LANGUAGE MODELS CANNOT TELL WHEN A DEFINITION IS MISSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contracts, technical documentation, and online communities give familiar words meanings of their own. When the defining text is missing from a language model's input, the model answers by the word's usual meaning, and nothing in its output signals the error. We study these silent misreadings on four corpora with diagnostic questions whose answers follow from a term's local definition. Frontier models answer 13% to 30% of these questions wrongly without the definition. The errors concentrate on familiar words given a new meaning, often repeat across ten samples, and persist across model generations, as 70% to 94% of the newest model's errors are also made by an older one. Sampling agreement and stated confidence rank them with AUROC 0.46 to 0.56 on redefined common words, and the calibrated confidence of a decision model reaches only 0.64 to 0.66 until the definition is supplied. Checking a small open model against the answer implied by the source ranks frontier models' errors with AUROC 0.72 to 0.85, where term frequency and rarity reach at most 0.63, and places two thirds of a contract library's estimated errors in a quarter of its questions. The source definition repairs most errors, whereas the model's own gloss of the term does not. Models can follow a definition they are given but cannot recognize when one is missing, so local definitions have to be located and supplied from the source.

1 INTRODUCTION

A distribution agreement in the Contract Understanding Atticus Dataset (CUAD) (Hendrycks et al., 2021) defines *Trademarks* as only the names listed in Exhibit A. Elsewhere it requires the distributor to report unauthorized use of the Trademarks. Suppose a contract assistant is asked whether the distributor must report a confusable name that is not in Exhibit A. A retriever returns the notification clause, which shares words with the question, but the definition sits in another section. In our CUAD retrieval experiments, the retrieved passages contain the needed definition for only 29% to 44% of questions across three retrievers. A model that reads the notification clause without the definition takes *trademarks* in its usual sense of any brand name and concludes that the name must be reported. GPT-5.6-terra gives exactly this answer without the definition and the correct one with it (Table 3). The answer sounds careful, and nothing in the retrieved text contradicts it. We call such an answer a *silent misreading*: a wrong answer about a locally defined term, given by the term's usual meaning, with no sign of doubt.

Such errors arise wherever language models read text written for a particular group. Contracts define capitalized terms, protocol documentation fixes the meaning of words such as *pool* or *position*, and online communities maintain glossaries of their own senses (Lucy & Bamman, 2021). These are also the settings in which organizations most often ask models about their own documents. The safeguards that deployed systems rely on, such as agreement across samples, confidence thresholds, and checks against retrieved evidence, all let a silent misreading pass. It also falls outside the categories used to evaluate model errors. Hallucination research separates factuality errors, which contradict world knowledge, from faithfulness errors, which contradict the input (Wang et al., 2025b). A silent misreading is neither, because the usual meaning is correct in general and the input says nothing against it. An error that neither runtime safeguards nor current benchmarks detect can persist in deployment unnoticed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The failure is also distinct from knowledge conflict (Longpre et al., 2021; Xu et al., 2024; Su et al., 2024; Zhao et al., 2026). That work places evidence in the context that contradicts what the model has learned and asks which source the model follows. In a silent misreading the contradicting evidence is absent. When it is supplied, models largely use it. On synthetic questions the source definition repairs 209 errors and harms only 6 of 532 correct answers, and on a human-written redefinition task two frontier models follow a stated meaning on 95% and 99.8% of items. The first problem is therefore to know which local definitions a model must be given, and only then whether it uses them.

We study this problem with diagnostic questions. For an occurrence of a defined term, a question tests a consequence of its local meaning, and its answer key comes from the source document. A small open model scores the keyed answer once per corpus, without any call to the model that will later read the text. We then ask what these scores reveal about frontier models. Four findings emerge.

1. **Silent misreadings are common, stable, and persist across model generations** (Section 5). Frontier models err on 13% to 30% of meaning-dependent questions without the definition. Newer models err less often but largely on the same occurrences, so upgrading the model reduces the problem without removing it.
2. **The model's own confidence does not reveal them** (Section 6). Sampling agreement, stated confidence, and the calibrated confidence of a decision model are close to chance on redefined common words, so runtime monitoring of the model's output cannot serve as the safeguard.
3. **A small model checked against the source finds them** (Section 7). A local 7B model ranks frontier models' errors about as well as another frontier model's answers do, while term frequency and rarity barely beat chance. It places two thirds of a contract library's estimated errors in under a quarter of its questions, so a corpus owner can audit a collection before deployment and reuse the audit as models change.
4. **Accurate local definitions repair them, and their content matters** (Section 8). The source definition repairs most errors, while the model's own gloss of the word does not. Local definitions must be written from the source rather than generated by the model, and term frequency is a reasonable default for deciding which to write first.

Together these findings identify a gap in current models and in how they are evaluated. Models can follow a local definition they are given, but they cannot tell when one is missing, and neither greater scale nor better calibration has closed this gap. The resulting errors fall outside the categories that benchmarks and runtime safeguards check, so they go unmeasured in exactly the settings where organizations deploy models on their own documents. Because the occurrences at risk are shared across models and can be located by a small model before deployment, where a general model's prior diverges from a community's usage becomes a measurable property of a text collection, much as term importance became measurable through inverse document frequency (Spärck Jones, 1972). Systems that read local text should therefore treat local definitions as data to be located and supplied from the source rather than as knowledge the model will infer (Section 9).

2 Related work

Hallucination, knowledge conflict, and context faithfulness. Hallucination research asks whether an answer agrees with world knowledge or with the supplied input (Wang et al., 2025b). Knowledge-conflict benchmarks place evidence in the context that contradicts parametric knowledge and measure which source the model follows (Longpre et al., 2021; Su et al., 2024; Zhao et al., 2026; Xu et al., 2024). Context-aware and adaptive contrastive decoding, together with steering methods, push the model toward the context (Shi et al., 2024; Wang et al., 2025a; Yuan et al., 2024; Anand et al., 2026; Li et al., 2025). Redefinition tasks state a new meaning and test whether the model adopts it (McKenzie et al., 2023; Stringli et al., 2025). All of these assume that the needed information is in the input or that the error contradicts it. Silent misreadings begin when the information is absent and the answer contradicts nothing the model can see. Contrastive decoding, which reweights information already in the input, cannot supply it. With Qwen2.5-14B-Instruct as a white-box reader, it lowers accuracy by 2.8 points $[-4.4, -1.0]$ on 1,200 CUAD questions when

Table 1: Three ways a missing definition can matter. The categories guide interpretation and are not labels assigned to the diagnostic items.

Situation	What is missing or misleading	Remedy
Meaning conflict	The usual reading supports a different answer from the local definition.	Supply the local definition with its scope.
Scope unspecified	The general concept fits, but its members or boundary in this document are unknown.	Supply the local binding. A dictionary gloss does not help.
Other information	The needed fact or rule is not part of the term's definition.	Retrieve the relevant evidence. More definitions do not help.

only retrieved passages are available (Appendix C). This work addresses the step before faithfulness, locating what a reader must be told.

Local language and word meaning. Studies of community language, semantic change, new concepts, and ordinary meaning in law describe how word senses vary and whether models know them (Lucy & Bamman, 2021; Tahmasebi et al., 2021; Schlechtweg et al., 2020; Mei et al., 2024; Purushothama et al., 2026). We measure how models behave on individual occurrences when a local meaning is missing and locate these occurrences before the model is called. A term-level measure of how far the local meaning is from the usual one separates designed term classes but predicts errors weakly on real corpora (Appendix B), because whether an occurrence is misread depends on what the question asks.

Error detection and question difficulty. Sampling consistency, semantic entropy, and verbalized confidence detect many hallucinations (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Tian et al., 2023). We identify a class of errors they miss. Item response theory and answer-plausibility methods estimate question difficulty across models (Vania et al., 2021; Byrd & Srivastava, 2022; Mozafari et al., 2026). Our score shares their premise that models agree on what is hard, and differs in being tied to a source occurrence and a local definition, so that it can be computed before deployment and paired with a specific remedy. SlimPLM uses a small model's answer to an incoming question to decide what a larger model should retrieve (Tan et al., 2024). Our proxy scores corpus questions before any request arrives, which complements retrieval and long-context use (Lewis et al., 2020; Liu et al., 2024).

3 SETTING

Local meaning and meaning-dependent questions. A *local meaning* is the meaning or scope that a document or community assigns to a term. A *reader* is a model that receives text and a question and answers it, and its *default reading* is its behavior when the local definition is absent. A *meaning-dependent question* asks about one occurrence of a term, and its correct answer follows from the local definition together with the occurrence. A wrong answer can arise in three ways (Table 1). The usual meaning can conflict with the local one, the local scope can be unknown although the usual concept fits, or the needed information can lie outside the definition. A risk score that ranks all three is useful for triage, and interventions with and without the definition show which errors a definition fixes.

Corpora. We use four corpora with different sources of local meaning. A synthetic collection of invented trading-desk and restaurant vocabulary controls which terms are redefined common words, which keep their usual meaning, and which are coinages. CUAD contains 510 commercial contracts with explicit definition clauses (Hendrycks et al., 2021). Reddit contains 35 communities whose members maintain glossaries of community terms (Lucy & Bamman, 2021). DeFi contains developer documentation for a decentralized exchange protocol. Diagnostic questions are four-option multiple-choice items drafted by a model from the source. The synthetic drafter saw only usage sentences, while the Reddit and DeFi drafters also saw the source definition. Appendix A gives counts and construction rules, and Table 8 lists the definition text used in each setting.

The synthetic and CUAD sources fully specify each answer, since the synthetic world is written from explicit specifications and contracts state their definitions. Reddit and DeFi answers often also rest on community practice or on documentation beyond the preserved excerpt. When two models of different families answered 20 hashed questions per corpus from the preserved source alone, both supported the stored key for 18 synthetic and 13 CUAD questions and for 3 Reddit and 5 DeFi questions (Appendix A). The main estimates therefore come from synthetic and CUAD questions, and Reddit and DeFi test how far the findings extend.

Readers. The readers are `gpt-5.6-terra`, `gpt-6-astra`, and DeepSeek Flash (V4 for synthetic questions and V4.1 for later runs). Jev (Almeida, 2026) is a decision model that returns a probability for each option together with a native confidence, and we use it to test whether a confidence signal built for this purpose detects the errors. Proxies are open Qwen2.5 models from 0.5B to 14B parameters and Llama-3.1-8B (Qwen Team, 2024; Grattafiori et al., 2024), run locally. Appendix F gives versions and decoding settings.

4 A RISK INDEX COMPUTED OFFLINE

The index assigns one score to each diagnostic question. Let z be a question with its source occurrence and keyed answer a^*, and let $p_S^0(a^* \mid z)$ be the probability that a proxy model S assigns to the keyed option when the definition is absent. The risk score is

$$R_S(z) = 1 - p_S^0(a^* \mid z). \tag{1}$$

A high R_S means that the proxy's default reading gives little support to the answer the source implies. R_S requires no call to the deployed model and is computed once per corpus. It does need an answer key, which the corpus owner can derive from the full source when the index is built, whereas a deployed model sees only the retrieved passages. The index therefore moves the comparison with the source from answer time to indexing time. We evaluate R_S by how well it ranks another model's errors on the same questions.

A second quantity evaluates a definition text rather than a question. With a definition text d prepended,

$$\Delta p_S(z) = p_S^d(a^* \mid z) - p_S^0(a^* \mid z) \tag{2}$$

measures how far that wording moves the proxy toward the source answer. Since $\Delta p_S = R_S + p_S^d(a^* \mid z) - 1 \leq R_S$, it adds information only about the definition. We use it in Section 8 to test whether a given wording repairs the errors it is meant to repair. Rewriting a definition changes Δp_S, so the score belongs to one definition text. Confidence intervals resample contracts for CUAD, communities for Reddit, and term records otherwise (Dror et al., 2018).

5 SILENT MISREADINGS ARE COMMON, STABLE, AND PERSIST ACROSS GENERATIONS

Table 2 reports how often readers misread. Without the definition, DeepSeek V4 Flash errs on 28.5% of synthetic questions and GPT-6-astra on 18.2%. In the contract library, GPT-5.6-terra errs on an estimated 13.2% of the 6,529 scored questions after reweighting the stratified sample, and the other corpora give 14.6% to 30.1%. With the definition text, error falls in every row, so these errors reflect missing information rather than an inability to answer the question.

The errors concentrate on familiar words given a new meaning. On synthetic terms whose local meaning conflicts with the usual one, DeepSeek V4 Flash errs on 44.2% of questions, against 25.3% on coined words and 14.8% on words that keep their usual meaning. GPT-6-astra shows the same order, with 36.5%, 12.7%, and 4.2% (Figure 1a). A familiar word misleads more than an unfamiliar one because the reader already has a confident reading of it. Table 3 shows three contract cases in which the usual meaning suggests a broader reading than the contract allows.

The errors are also stable. For 761 synthetic questions, DeepSeek V4 Flash gave a greedy answer and ten sampled answers at temperature 1.0. Of the 229 greedy errors, 93 (40.6%) repeat the same wrong answer in all ten samples, and the majority answer is wrong for 142 (62%). Majority voting would correct 38% of greedy errors, but the errors that survive concentrate on redefined common

Table 2: Reader error without and with the definition text, and AUROC for ranking errors without the definition. R is the proxy risk score. Freq. is the better of term frequency and term rarity as a ranking score. Δp ranks the same errors by the change a definition makes, which is not its intended use (Section 4). The definition text is a model-drafted source-based gloss for synthetic, Reddit, and DeFi, and extracted contract wording for CUAD. Rows differ in reader and question construction (Appendix A). CUAD is risk-stratified.

Corpus	Reader	n	Error, no def.	Error, def.	R AUROC [95% CI]	Freq.	Δp
Synthetic	DeepSeek V4 Flash	456	0.285	0.029	0.848 [0.811, 0.883]	0.540	0.745
Synthetic	GPT-6-astra	456	0.182	0.018	0.809 [0.763, 0.856]	0.629	0.656
Reddit	DeepSeek V4.1 Flash	1,391	0.301	0.185	0.808 [0.785, 0.835]	0.544	0.520
Reddit	GPT-6-astra	1,391	0.146	0.119	0.720 [0.676, 0.761]	0.588	0.429
DeFi	DeepSeek V4.1 Flash	379	0.282	0.193	0.809 [0.763, 0.856]	0.510	0.458
DeFi	GPT-5.6-terra	379	0.230	0.148	0.792 [0.742, 0.839]	0.589	0.380
DeFi	GPT-6-astra	379	0.169	0.135	0.764 [0.704, 0.815]	0.603	0.342
CUAD	GPT-5.6-terra	1,500	0.213	0.160	0.815 [0.789, 0.840]	0.507	0.464

Table 3: Contract questions from the validation sample that GPT-5.6-terra answers wrongly without the definition and correctly with it. The usual meaning of each term suggests a broader reading. Percentiles are within the 6,529-question risk index.

Term	Local definition	Question (abridged)	Answer implied by the definition	Risk
Products	dietary supplements manufactured within the United States	Which products must the endorser have used to give honest findings?	only U.S.-made supplements	97th
Trademarks	the trademarks and trade names identified in Exhibit A	Must the distributor report a confusable third-party name not in Exhibit A?	report only names listed in Exhibit A	99.7th
Term	the period from the Signing Date to the End Date	During which period must the required insurance stay in force?	from signing to the stated end date, not until termination	98th

words, where 65% survive against 43% on usual-meaning terms. A system that relies on agreement across samples (Wang et al., 2023) removes the unstable errors and keeps the ones tied to local meaning.

Newer models do not remove the problem. In every corpus with several readers, the newest reader makes the fewest errors, but its errors are largely a subset of older readers' errors. On questions answered by every reader of a corpus, 78% of GPT-6-astra's synthetic errors are also errors of DeepSeek V4 Flash, 70% of its Reddit errors are also errors of DeepSeek V4.1 Flash, and 94% of its DeFi errors are also made by DeepSeek V4.1 Flash or GPT-5.6-terra. If errors were independent, these shares would be 29%, 30%, and 36% (Figure 1b). Model improvement shrinks the set of misread occurrences without moving it, so the occurrences at risk depend on the corpus as much as on the model.

6 THE MODEL'S OWN CONFIDENCE DOES NOT REVEAL THE ERRORS

If misreadings came with uncertainty, the detectors already used to flag hallucinations would find them. We test three signals from the reader's own samples on the 456 synthetic questions shared with the proxy: disagreement among ten sampled answers, entropy of the sampled option distribution, and negative stated confidence (Manakul et al., 2023; Kuhn et al., 2023; Tian et al., 2023). None of them uses the answer key. For DeepSeek V4 Flash errors they reach AUROC 0.616 $[0.579, 0.656]$, 0.616 $[0.579, 0.656]$, and 0.617 $[0.558, 0.681]$ (Figure 2a). On redefined common words they fall to 0.557, 0.557, and 0.463, against 0.60 to 0.72 on invented and usual-meaning terms (Appendix A). The detectors fail precisely where local meaning is at stake. Our option-level entropy is narrower than semantic entropy over free-form answers (Farquhar et al., 2024), but the result agrees with

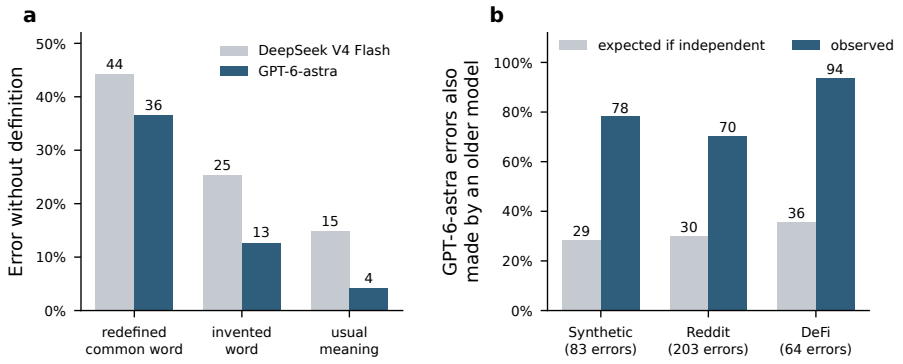

Figure 1: (a) Error without the definition on synthetic questions by designed term class. Familiar words given a new meaning are misread most often. (b) Share of GPT-6-astra's errors that an older model also makes on the same questions, against the share expected if errors were independent, which is the older models' overall error rate on those questions (for DeFi, the rate at which DeepSeek V4.1 Flash or GPT-5.6-terra errs).

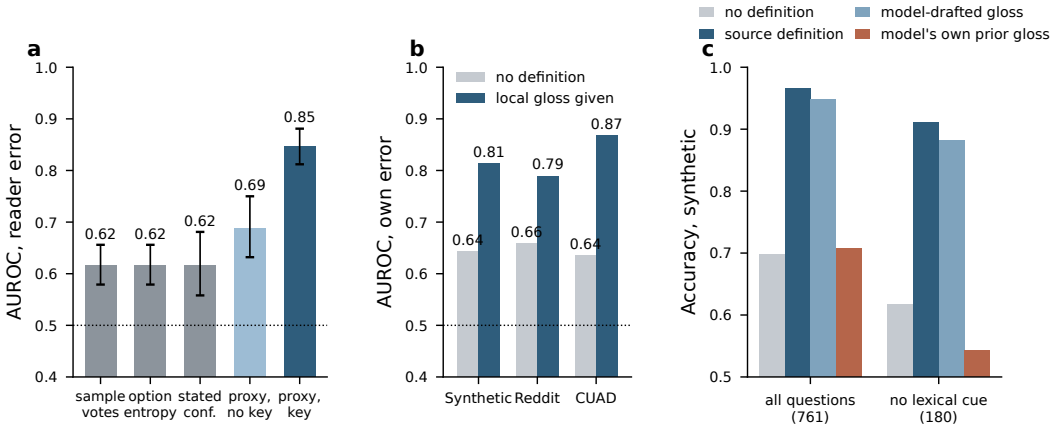

Figure 2: (a) AUROC for ranking DeepSeek V4 Flash errors on 456 synthetic questions, with 95% intervals from resampling terms. Gray bars are key-free signals from the reader's ten samples and stated confidence. The light bar is the key-free uncertainty of the 7B proxy (one minus its largest option probability). The dark bar is the proxy score R, which checks the proxy against the source-implied answer. The dotted line marks chance. (b) Jev's native confidence as a detector of its own errors, without a definition and with a model-drafted local gloss. (c) DeepSeek V4 Flash accuracy on synthetic questions under four conditions: no definition, the source definition, a model-drafted gloss, and the model's own gloss of the term written under a domain prompt. The right group keeps only questions where no option is singled out by word overlap with any of the three texts.

the stability above. The small model's own uncertainty is no better. One minus its largest option probability reaches only 0.537 and 0.523 on redefined common words for the two readers.

Better calibration does not close the gap. Jev (Almeida, 2026) is a decision model built to report calibrated confidence. Its confidence ranks its own errors without a definition with AUROC 0.643 on synthetic, 0.660 on Reddit, and 0.636 on CUAD questions, and with a model-drafted local gloss added, with 0.815, 0.790, and 0.868 (Figure 2b). The same confidence channel works once the local meaning is present. The limitation is therefore one of information rather than calibration. A model cannot flag a meaning it does not know has changed, so confidence thresholds, abstention rules, and sample agreement tuned on ordinary errors pass these answers.

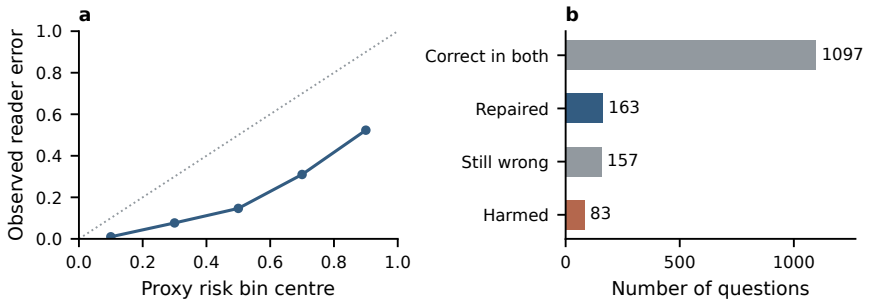

Figure 3: CUAD library validation with 300 questions per proxy risk bin and GPT-5.6-terra as reader. (a) Observed error rises with proxy risk. The dotted line marks equality, which the score does not attain because it ranks rather than estimates error. (b) Outcomes when the extracted contract definition is supplied. Repairs outnumber harms, and about half of the original errors remain.

7 A SMALL MODEL CHECKED AGAINST THE SOURCE FINDS WHERE FRONTIER MODELS FAIL

Cross-reader prediction. The risk score R compares the proxy's default reading with what the source implies, instead of asking whether the proxy is uncertain. The proxy never sees the reader's answers, yet its score ranks them. On the 456 synthetic questions, R from Qwen2.5-7B ranks DeepSeek V4 Flash and GPT-6-astra errors with AUROC 0.848 and 0.809, and on the 1,500 CUAD validation questions it ranks GPT-5.6-terra errors with 0.815 (Table 2). On Reddit and DeFi it reaches 0.720 to 0.809. This is comparable to using another frontier model's observed answers on the same questions, which give 0.640 to 0.892 (Appendix A), and it needs no call to any frontier model. Term frequency and rarity, the statistics a corpus index built without questions would use, reach at most 0.63 in any setting (Table 2). Finding these errors requires a check against the source, not a property of the term alone. The signal grows with proxy size, from 0.699 at 0.5B parameters to 0.848 at 7B and 0.867 at 14B on synthetic questions, so the index can be built on a laptop and improves with larger open models.

Risk stratification across a contract library. We score 6,529 diagnostic questions from 303 CUAD contracts and ask GPT-5.6-terra 300 questions from each of five risk bins. Observed error rises monotonically from 1.0% in the lowest bin to 52.3% in the highest (Figure 3a). Reweighted to the full library, error is 13.2% and AUROC is 0.846. The two highest bins hold 23% of the questions and 66% of the estimated errors, so a reviewer who checks only these bins sees two thirds of the errors. Because the index is built when the documents are collected, this review can happen before any request arrives, and the cross-generation overlap in Section 5 indicates that the ranking remains relevant as the deployed model changes.

What the score measures. R ranks errors of every kind in Table 1. On synthetic terms that keep their usual meaning its AUROC is 0.921 for DeepSeek V4 Flash and 0.846 for GPT-6-astra, based on 21 and 6 errors, and on Reddit control words it is 0.842 and 0.815 (Appendix A). Part of what R captures is therefore difficulty that models share on meaning-dependent questions. R is thus a general index of where a reader's default reading departs from the source, and Section 8 shows which of the flagged errors a definition repairs.

Robustness to question quality. Model-drafted keys can be wrong. An author answered 27 questions from the source before seeing the stored keys, chosen as the 17 on which the two screening models disagreed and 10 stratified controls. After reconciliation with the sources, 10 keys are supported, 14 questions are invalid or not answerable from the preserved source, and 3 are borderline (Appendix A). Removing the 14 invalid questions changes the AUROC of R by at most 0.002 in all eight reader settings and the CUAD library AUROC from 0.846 to 0.847, and leaves the synthetic count of 93 stable errors unchanged. The estimates above do not depend on the defective questions that were found.

Table 4: Predicting which of a reader's errors the definition text repairs, as AUROC for ranking repaired above unrepaired errors among errors made without the definition. R and Δp use the same diagnostic question in the middle columns. The last column uses other scored questions of the same term record, with the available failed-question count in parentheses. CUAD uses the full proxy-scored frame for sibling estimates. Other rows use their available scored siblings. The CUAD 95% interval is $[0.494, 0.661]$.

Setting	Bare failures	R	Δp	Other-item Δp (n)
Synthetic / DeepSeek V4 Flash	130	0.445	0.983	0.914 (130)
Synthetic / GPT-6-astra	83	0.421	0.962	0.976 (83)
Reddit / DeepSeek V4.1 Flash	419	0.453	0.858	0.824 (418)
Reddit / GPT-6-astra	203	0.554	0.812	0.864 (203)
DeFi / DeepSeek V4.1 Flash	107	0.501	0.863	0.850 (107)
DeFi / GPT-5.6-terra	87	0.568	0.794	0.799 (87)
DeFi / GPT-6-astra	64	0.476	0.846	0.874 (64)
CUAD / GPT-5.6-terra	320	0.462	0.712	0.581 (190)

8 Accurate local definitions repair the errors, and their content matters

The repair comes from the source, not from the word. On 761 synthetic questions whose drafter saw usage sentences but no definition, DeepSeek V4 Flash accuracy is 0.699 without a note, 0.966 with the source definition, 0.949 with a model-drafted gloss written from that source, and 0.708 with the reader's own gloss of the term written under a domain prompt (Figure 2c). The source definition repairs 209 errors and harms 6 correct answers, while the reader's own gloss repairs 86 and harms 79. On the 180 questions where no option is singled out by word overlap with any of the three texts, accuracy still rises from 0.617 to 0.911 with the source definition, against 0.544 with the reader's gloss (Appendix C). Asking a model to state its understanding of a term before answering therefore restates its prior and does not help. The repair comes from local content.

Readers also follow a new meaning when it is stated outright. On 1,244 human-written redefinition items of the Inverse Scaling task (McKenzie et al., 2023), GPT-5.6-terra and GPT-6-astra answer by the stated meaning on 95.0% and 99.8% of items, and with the redefinition removed both answer by the usual meaning on about 98%. DeepSeek V4.1 Flash follows the stated meaning on only 63.0%. Disregarding a stated definition is thus a separate failure whose rate differs sharply between readers. For the two GPT readers, nearly all of the failure lies in not having been told.

The wording of a definition sets both repairs and harms. The comparison extends to Reddit and DeFi (Table 17 in Appendix C). On Reddit glossary questions, a model-drafted gloss written from the community's entry repairs 171 errors and harms 58 correct answers, while the community entry itself repairs 243 and harms 9. On DeFi the counts are 42 and 20 against 70 and 3, and the reader's own gloss repairs 82 and harms 117 on Reddit and repairs 40 and harms 23 on DeFi. An inaccurate or loosely scoped note can turn a correct answer into a wrong one. Δp detects this in advance for a given wording, and here the two scores divide the work. Among a reader's errors without a definition, R cannot tell which ones a definition will repair, with AUROC 0.42 to 0.57, while Δp ranks repaired errors above unrepaired ones with 0.712 to 0.983 across eight settings (Table 4). The effect also carries over within a term. Δp averaged over the other questions of the same term predicts repair on a held-out question with 0.581 to 0.976, so a wording tested on a few uses of a term indicates how it will act on the rest. R locates the errors, and Δp tells whether a given definition fixes them. A definition can therefore be tested on the diagnostic questions before it is deployed.

A definition is not always sufficient. In CUAD, extracted contract definitions repair 163 errors and harm 83 correct answers, raising accuracy from 0.787 to 0.840 (Figure 3b). Of 320 original errors, 157 remain wrong. Contract definitions often refer to schedules or other clauses, so a system has to retrieve the linked evidence as well as the definition itself.

Choosing which definitions to write first. In a 215-term Reddit comparison at a budget of 25% of total definition words, a library chosen by term frequency is 3.5 and 4.5 points more accurate for the two readers than one chosen by frequency times mean Δp on the terms where the two policies differ, and a small CUAD comparison also does not favor Δp (Appendix E). Neither difference is significant. Ordering entries by frequency is a reasonable default, and the risk index then shows where readers still fail.

Public benchmarks rarely test these boundaries. On public human-written tasks, adding definitions changes scores little. CUAD clause extraction changes by at most 0.3 F1 points and ContractNLI (Koreeda & Manning, 2021) accuracy by at most 1.1 points (Appendix D). Both tasks ask what a clause says rather than whether a particular case falls inside a defined term, so few of their items turn on a definition's boundary. The same contracts produce errors on 13.2% of the diagnostic questions that do. Silent misreadings concentrate on boundary cases, and evaluating systems that read local text requires questions built to reach them.

9 Implications of the findings

A class of error that current evaluations do not measure. Error taxonomies for language models ask whether an answer contradicts the world or contradicts the input (Wang et al., 2025b). Silent misreadings do neither. The input is consistent with the answer but incomplete relative to the document it came from, and the answer is wrong only with respect to that document. Such errors become visible only when an answer is compared with the full source rather than with the supplied input, and the small effect on public benchmarks indicates that this comparison is rarely made rather than that the errors are rare. Evaluations of systems that read local text, including retrieval-augmented generation, need questions that reach the boundaries of local definitions.

Knowing that a definition is needed is a distinct capability. Models largely know what they know (Kadavath et al., 2022), and calibration and selective prediction exploit this to abstain when uncertain (Guo et al., 2017; Geifman & El-Yaniv, 2017). A redefined common word defeats this mechanism, because the model is not uncertain about the word. It is confident in the wrong sense. The strongest readers follow a stated definition on nearly every item and misread largely the same occurrences as their predecessors, so the remaining failure is not a lack of capacity to use local meaning. It is the absence of any internal signal that local meaning is needed. Recognizing that a familiar word may carry a local meaning is thus a capability that neither scale nor calibration has produced, and a natural target for evaluation and training.

Where a general model's prior diverges from a community's usage is a property of the corpus. Because the occurrences at risk are shared across model generations and ranked by a small model checked against the source, where a general reader can be trusted on local text becomes a measurable property of a collection. Like inverse document frequency for retrieval (Spärck Jones, 1972), the risk index is computed once over a collection and indicates where every subsequent reader is likely to need the collection's own definitions. It also shows that a weaker model can locate where a stronger one fails, as in weak-to-strong settings (Burns et al., 2023). The same comparison applies wherever text is written for a group whose vocabulary differs from general usage, as in software engineering terminology (Ejaz et al., 2026), community slang (Sun et al., 2024), and organizational documents.

10 Limitations and conclusion

The diagnostic questions are drafted by models and are forced-choice. They show that errors occur and can be ranked, but they are not requests written by users, and the gains from definitions are measured on these questions. The Reddit and DeFi answers more often rest on material beyond the preserved source, which is why the synthetic and CUAD questions carry the main estimates. The risk score is computed for corpus questions, and its use for routing incoming requests remains to be shown. The selection experiments are small and compare frequency with Δp-based choice only. Within these limits, in the corpora and models we test, frontier models misread locally defined terms often, stably, and without signaling it. A small model checked against the source locates these errors in advance, and definitions written from the source repair them.

REPRODUCIBILITY STATEMENT

We release a risk index over 6,529 retained CUAD diagnostic questions, 1,500 of them with frontier-reader outcomes, together with question sets, reader outcomes, proxy scores, and an experimental command-line pipeline for new corpora. The GitHub repository includes model records, analysis scripts, and verification of the main numerical results `https://anonymous.4open.science/r/llm-local-meaning-diagnostics-FDF7/`. The appendix specifies each evaluation unit and the limits of the source review.

AI USE STATEMENT

We used generative AI tools to help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement code, assist with translation, clean and reformat datasets, interpret results, and support qualitative and thematic data analysis. We have not used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Diogo Almeida. Introducing system one models & Jev, September 2026. URL `https://typesafe.ai/blog/introducing-system-one-models-and-jev`. TypeSafe AI company announcement, September 15, 2026. Accessed September 26, 2026.

Nikhil Anand, Shwetha Somasundaram, Anirudh Phukan, Apoorv Saxena, and Koyel Mukherjee. ContextFocus: Activation steering for contextual faithfulness in large language models. *arXiv preprint arXiv:2601.04131*, 2026.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision, 2023. URL `https://arxiv.org/abs/2312.09390`.

Matthew Byrd and Shashank Srivastava. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 119–130, 2022.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pp. 2318–2335, 2024.

DeepSeek. Introducing DeepSeek-V4.1-Flash, September 2026. URL `https://www.deepseek.com/en/news/deepseek-v4-1-flash/`. Official model and API announcement, September 10, 2026. Accessed September 26, 2026.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392, 2018.

Huzaifa Ejaz, Fabian C. Peña, and Steffen Herbold. Large language models have unreliable understanding of software engineering terminology, 2026. URL `https://arxiv.org/abs/2607.06004`.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017. URL `https://arxiv.org/abs/1705.08500`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. URL https://arxiv.org/abs/1706.04599.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. *Advances in neural information processing systems*, 36:47934–47959, 2023.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/6ea9ab1baa0efb9e19094440c317e21b-Paper-round1.pdf.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Yuta Koreeda and Christopher D Manning. Contractnli: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919, 2021.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0cEZyhHEks.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7052–7063, 2021.

Li Lucy and David Bamman. Characterizing english variation across social media communities with bert. *Transactions of the Association for Computational Linguistics*, 9:538–556, 2021.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

IR McKenzie, A Lyzhov, M Pieler, A Parrish, A Mueller, A Prabhu, E McLean, A Kirtland, A Ross, A Liu, A Gritsevskiy, D Wurgaft, D Kauffman, G Recchia, J Liu, J Cavanagh, M Weiss, S Huang, The Floating Droid, T Tseng, T Korbak, X Shen, Y Zhang, Z Zhou, N Kim, SR Bowman, and E Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=DwgRm72GQF.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. SLANG: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12558–12575, 2024.

Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. Question difficulty estimation for large language models via answer plausibility scoring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11124–11151, 2026.

OpenAI. GPT-5.6 Terra model, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra. Official API model specification; knowledge cutoff February 16, 2026. Accessed September 26, 2026.

OpenAI. GPT-6 Astra model, 2026b. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Official API model specification; knowledge cutoff April 30, 2026. Accessed September 26, 2026.

Abhishek Purushothama, Junghyun Min, Brandon Waldon, and Nathan Schneider. Prompting from the bench: Large-scale pretraining is not sufficient to prepare LLMs for ordinary meaning analysis. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pp. 4552–4580, 2026.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the fourteenth workshop on semantic evaluation*, pp. 1–23, 2020.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526.

Elena Stringli, Maria Lymperaiou, Giorgos Filandrianos, Athanasios Voulodimos, and Giorgos Stamou. Pitfalls of scale: Investigating the inverse task of redefinition in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9445–9469, 2025.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. *Advances in Neural Information Processing Systems*, 37:103242–103268, 2024.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. Toward informal language processing: Knowledge of slang in large language models, 2024. URL https://arxiv.org/abs/2404.02323.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen (eds.), *Computational approaches to semantic change*, pp. 1–91. Language Science Press, Berlin, 2021. doi: 10.5281/zenodo.5040302.

Jiejun Tan, Zhicheng Dou, Yutao Zhu, Peidong Guo, Kun Fang, and Ji-Rong Wen. Small models, big insights: Leveraging slim proxy models to decide when and what to retrieve for llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4420–4436, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Uniswap Labs. Unichain mainnet is here: A new home for DeFi, February 2025a. URL https://blog.uniswap.org/unichain-mainnet-is-here. February 11, 2025. Accessed September 26, 2026.

Uniswap Labs. Uniswap v4 is here: A new era of DeFi, January 2025b. URL https://blog.uniswap.org/uniswap-v4-is-here. January 31, 2025. Accessed September 26, 2026.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel Bowman. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1141–1158, 2021.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11636–11652, 2025a.

Pengbo Wang, Chaozhuo Li, Chenxu Wang, Liwen Zheng, Litian Zhang, and Xi Zhang. One SPACE to rule them all: Jointly mitigating factuality and faithfulness hallucinations in LLMs. *Advances in Neural Information Processing Systems*, 38:157809–157828, 2025b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. arXiv:2203.11171.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3903–3922, 2024.

Tianzhe Zhao, Jiaoyan Chen, Shuxiu Zhang, Haiping Zhu, Qika Lin, and Jun Liu. Exploring knowledge conflicts for faithful LLM reasoning: Benchmark and method. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2464–2474, 2026. URL https://doi.org/10.1145/3805712.3809650.

A DATA AND QUESTION CONSTRUCTION

Diagnostic questions are generated from source passages and local definitions. A question specifies an option set and a source-grounded answer key. These keys support behavioral measurement but do not constitute expert adjudication of every application scenario. The supplementary inventory distinguishes question-level results from term-level means and keeps separate task versions separate.

Synthetic vocabulary. The source design uses invented trading-desk and restaurant-group vocabulary. Categories and local definitions are specified before text generation, and the diagnostic intervention uses a model-drafted gloss derived from the source. The archived uncertainty experiment contains 763 questions from 197 nonempty term records. Excluding two questions affected by a source-definition conflict leaves 761 for analysis. The archived 7B proxy overlap has 457 questions from 119 records. Excluding its one affected question leaves 456. The usage-based inline experiment contains 965 questions. Frequency is deliberately varied in the synthetic design, so rarity comparisons concern that design rather than a natural corpus distribution.

CUAD. The source corpus contains 510 contracts. Extraction retains 2,174 term records in 303 contracts and identifies 24,880 candidate occurrences. Question drafting and per-term caps yield a 6,529-question scored frame. The stratified validation sample contains 1,500 questions in 1,012 term records and 215 contracts. A separate definition-anchored task contains 671 questions in 187 terms. These task sets have different construction rules and are analyzed separately.

The introductory example is record `cuadfull:379:Trademarks`, question at context index 2. Its definition restricts Trademarks to the trademarks and trade names identified in Exhibit A. The usage requires the distributor to notify the manufacturer of unauthorized use of the Trademarks, so the notification duty covers only the listed names.

Reddit. The data contain 393 term records, of which 352 have questions. The 1,391 retained questions span 35 communities. The RAG workload contains 851 glossary-word questions and 540 control-word questions. All 393 stored diagnostic `gloss_used` fields are marked as model drafted. Among the 218 glossary terms with a source entry, none of these drafted texts equals `gloss_gold`. Controls need not have a source entry. A comparison involving `gloss_used` therefore measures that rewrite, not the community source entry alone.

DeFi. The set contains 379 questions from 97 nonempty term records. It tests technical vocabulary in DeFi documentation. All 98 stored candidate glosses are model drafted and differ from the recorded source glosses. Official model cards give knowledge cutoffs of February 16, 2026 for `gpt-5.6-terra` and April 30, 2026 for `gpt-6-astra` (OpenAI, 2026a;b). The collected documentation snapshot contains later file dates, but each term links a median of 18 documents and questions have no verified first-publication date for their answer-bearing rule. Uniswap v4 and Unichain were public in 2025 (Uniswap Labs, 2025b;a). Thus the GPT readers' R AUROCs of 0.792 and 0.764 measure cross-reader error ranking on these questions. They do not rule out knowledge of earlier documentation or prove that answers were not memorized. The analysis does not depend on an assumption that this material was absent from model training.

Question and answer-key validity. A fixed SHA-256 rule selected 20 historical questions from each corpus before comparing any new answers with the stored keys. Reddit was split evenly between its ordinary question frame and a frozen strategy-disagreement frame. This 10/10 split is not a natural prevalence estimate. DeepSeek V4.1 Flash and Claude Sonnet 5 independently read the same source cards and question-only excerpts from linked documents. Both were run without explicit reasoning. DeepSeek used temperature zero and a 550-token output cap. Sonnet 5 rejected temperature zero, so its host used the provider default and a 1,000-token cap. All 80 Sonnet responses ended normally, but its host did not expose provider response IDs. The models saw neither the stored key nor reader outcomes. Long contracts and DeFi documents were excerpted, and the Reddit archive preserves glossary entries and selected comments rather than every complete original post.

Human source review. From the 80 fixed-hash questions, an author answered 27 from the available source before seeing the stored key. Seventeen were model-screen disagreement cases. Ten

Table 5: Model source screening on 20 questions per corpus. "Both key" means both models gave a unique source-supported answer matching the stored key. "Both other" means they agreed on another answer. "Split answer" means two supported but different answers. "One supported" means one model judged the source sufficient and the other did not. "Neither supported" includes insufficient or ambiguous evidence. These are screening categories, not expert answer-key error rates.

Corpus	Both key	Both other	Split answer	One supported	Neither supported
Synthetic	18	0	0	1	1
CUAD	13	1	1	4	1
Reddit	3	0	0	3	14
DeFi	5	0	0	7	8

were sampled within corpus and screen categories. The author selected an option on 11 and marked 16 undetermined. A different letter was not automatically counted as a key error: the local definition supports the stored key in some such cases. Subsequent source reconciliation, conducted after viewing the key, produced Table 6. It is not an independent blinded second annotation and yields no agreement coefficient. The reviewed defects include a redacted exhibit needed to determine a

Table 6: Source review of 27 selected diagnostic questions: 17 model-screen disagreements and ten stratified controls. Counts describe this enriched sample, not corpus-wide defect rates.

Corpus	Key supported	Invalid or source-insufficient	Borderline	Reviewed
Synthetic	3	1	0	4
CUAD	2	6	1	9
Reddit	3	1	1	5
DeFi	2	6	1	9
Total	10	14	3	27

manufacturer's duty, an item with literal A/B/C/D placeholders as its options, a mandatory-removal question where the contract permits removal, and a blocked-swap status with two possible causes. In one CUAD item the stored option names user guides and technical manuals as rights not granted, although the contract explicitly grants rights to Documentation. The item was removed rather than rekeyed using probabilities computed for the old answer. Conversely, the community entries for *Fill*, *Burst*, and *Standard* can resolve constructed Reddit scenarios without an archived post describing that exact scenario. These distinctions require source reading rather than equating an author-model disagreement with a wrong key. Removing the 14 clearly invalid or source-insufficient reviewed

Table 7: Deletion sensitivity after removing the 14 reviewed questions classified as clearly invalid or source-insufficient. Other questions retain their historical keys. This is a local sensitivity analysis, not a validated-only re-estimate. Δp AUROC targets repair among observed bare errors on the same question.

Setting	Items retained	R error AUROC	Conditional repair Δp AUROC
Synthetic / DeepSeek V4 Flash	455/456	.848 → .847	.983 → .983
Synthetic / GPT-6-astra	455/456	.809 → .809	.962 → .962
Reddit / DeepSeek V4.1 Flash	1,390/1,391	.808 → .808	.858 → .860
Reddit / GPT-6-astra	1,390/1,391	.720 → .720	.812 → .812
DeFi / DeepSeek V4.1 Flash	373/379	.809 → .808	.863 → .863
DeFi / GPT-5.6-terra	373/379	.792 → .791	.794 → .794
DeFi / GPT-6-astra	373/379	.764 → .764	.846 → .852
CUAD / GPT-5.6-terra	1,494/1,500	.815 → .817	.712 → .714

items changes R error-ranking AUROC by at most 0.00203 and same-question repair Δp AUROC by at most 0.00601 across the eight matched settings. CUAD design-weighted R AUROC changes from 0.8456 to 0.8467, and the corrected synthetic repeated-sampling count remains 93 identical wrong answers among 229 greedy errors. Removing the three borderline items as well has little effect on these point estimates. The 17 targeted items and ten stratified controls are not a random draw

from all historical questions. Their category counts are not defect rates. Most historical questions remain unreviewed, and this local deletion check cannot validate them. The 24 items in the original 80 for which neither screening model found a source-supported unique answer remain unresolved rather than being treated as wrong keys.

Definition version by CUAD task. The 1,500-question CUAD library builds its proxy and reader definition arm from `def_raw`, the extracted contract wording. The distinct 671-question definition-anchored CUAD package instead has an automatically drafted `glossC` arm and a separate original-contract `glossG` arm. All 204 small-package term records mark `gloss_used` as drafted and different from the source entry. Archived per-question prompts confirm that all 671 `glossC` notes match the drafted text and all 671 `glossG` notes match the original-contract text. These CUAD tasks must not be described as one common definition intervention.

Eight nonexchangeable reader settings. Table 8 records the reader alias, intervention text, and information visible to the question drafter. The synthetic questions were drafted from usage contexts without the source definition in the drafting prompt. Reddit and DeFi questions were drafted with their source entries, which also informed answer keys. The later reader intervention supplied a different model-drafted gloss. CUAD questions were drafted from an extracted contract definition and one usage occurrence. Within a corpus, additional readers answered the same diagnostic questions. Across corpora, the question construction and definition versions differ. The rows cannot be pooled as replications of one fixed data-generating process.

Table 8: Reader settings in Table 2. Synthetic DeepSeek calls used V4 Flash. Later calls used V4.1 Flash. DeepSeek documents the API IDs `deepseek-v4-flash` and `deepseek-flash`, respectively. "Source entry" describes drafting input and does not imply that the reader's `glossC` arm used that entry.

Corpus	Reader API alias	Definition given to reader	Question drafting input
Synthetic	DeepSeek V4 Flash	Model-drafted `glossC`	Invented-term usage contexts. No source definition in drafting prompt
Synthetic	`gpt-6-astra`	Model-drafted `glossC`	Same synthetic questions
Reddit	DeepSeek V4.1 Flash	Model-drafted `glossC`	Community glossary entry and comment contexts
Reddit	`gpt-6-astra`	Model-drafted `glossC`	Same Reddit questions
DeFi	DeepSeek V4.1 Flash	Model-drafted `glossC`	Source glossary entry and documentation contexts
DeFi	`gpt-5.6-terra`	Model-drafted `glossC`	Same DeFi questions
DeFi	`gpt-6-astra`	Model-drafted `glossC`	Same DeFi questions
CUAD	`gpt-5.6-terra`	Extracted contract `def_raw`	One definition and one contract usage per question

Designed-class control for mechanism specificity. The synthetic vocabulary assigns terms to prior-conflict, prior-consistent, and no-ordinary-prior classes before generating diagnostic questions. Table 9 evaluates R against reader bare errors separately in each class, with term-record bootstrap intervals. The score ranks errors in the prior-consistent class, where a conflicting usual meaning was not designed into the term. It also ranks errors for terms with no ordinary prior. These findings show that the observed error-ranking performance is not specific to meaning conflict. The classes are synthetic design labels, not independent diagnoses of each error, and the GPT-6-astra prior-consistent row has only six errors.

Table 9: Error-ranking R by designed synthetic term class. Reader cells give errors, AUROC [95% CI]. Both readers answer the same 456 questions. Intervals resample term records.

Class	Questions	DeepSeek V4 Flash	GPT-6-astra
Prior conflict	156	69, 0.803 [0.739, 0.864]	57, 0.777 [0.697, 0.854]
Prior consistent	142	21, 0.921 [0.870, 0.966]	6, 0.846 [0.710, 0.993]
No ordinary prior	158	40, 0.779 [0.704, 0.849]	20, 0.720 [0.629, 0.819]

Reddit non-glossary controls. The Reddit workload also contains 540 questions attached to terms outside the community glossary. DeepSeek V4.1 Flash has 0.235 bare-error rate and R AUROC 0.842 on these controls, versus 0.343 and 0.781 on 851 glossary-term questions. GPT-6-astra has corresponding control values 0.143 and 0.815, versus 0.148 and 0.659 on glossary terms. The non-glossary terms are a useful difficulty reference, but absence from the collected glossary does not prove absence of every local sense. The separate source audit also finds substantial Reddit question-answerability uncertainty.

Other readers as a difficulty reference. On identical questions, another reader's observed bare correctness can serve as a coarse indication of shared item difficulty. Table 10 compares that binary or two-reader error-count score with the continuous proxy R. The comparator uses another frontier model's answers and is therefore not a low-cost deployment baseline or a statistical upper bound. Its resolution differs from R, and both may respond to the same local-meaning failure. The proxy was not rescored with the source usage or target term removed. These measurements cannot isolate a semantic component of R.

Table 10: AUROC for predicting a target reader's bare errors on the same questions. "Other readers" is one observed reader error or the mean of two other reader errors. All models share the historical question and answer key.

Corpus	Target reader	n	Proxy R	Other readers
Synthetic	DeepSeek V4 Flash	456	.848	.722
Synthetic	GPT-6-astra	456	.809	.804
Reddit	DeepSeek V4.1 Flash	1,391	.808	.640
Reddit	GPT-6-astra	1,391	.720	.736
DeFi	DeepSeek V4.1 Flash	379	.809	.750
DeFi	GPT-5.6-terra	379	.792	.867
DeFi	GPT-6-astra	379	.764	.892

Stability and key-free signals by term class. Table 11 splits the synthetic stability and detection results by the designed term class. Stability columns use the 761-question sampling frame for DeepSeek V4 Flash. Detection columns use the 456 questions shared with the proxy, where error counts are smaller. Sample disagreement and option entropy give identical rankings on these ten-sample votes, so only disagreement is shown. The key-free proxy score is one minus the proxy's largest option probability, recomputed with the released scoring prompt. The rerun reproduces the stored keyed probabilities to within 0.005. Intervals resample term records.

Table 11: Synthetic stability and error-detection AUROC by term class for DeepSeek V4 Flash. "Same wrong" counts greedy errors repeated identically in all ten samples. "Majority wrong" counts errors that survive majority voting.

	Stability (761 questions)			Detection AUROC (456 questions)				
Class	Errors	Same wrong	Majority wrong	Errors	Disagreement	Stated conf.	Proxy, no key	Proxy R
Redefined common word	168/461	74 (44%)	110 (65%)	69	0.557 [0.492, 0.622]	0.463 [0.376, 0.554]	0.537 [0.439, 0.636]	0.803 [0.738, 0.864]
Coinage	40/158	12 (30%)	23 (58%)	40	0.641 [0.570, 0.720]	0.715 [0.637, 0.785]	0.718 [0.615, 0.816]	0.779 [0.704, 0.848]
Usual meaning	21/142	7 (33%)	9 (43%)	21	0.603 [0.517, 0.692]	0.644 [0.552, 0.730]	0.810 [0.710, 0.897]	0.921 [0.871, 0.965]
All	229/761	93 (41%)	142 (62%)	130	0.616 [0.579, 0.656]	0.617 [0.558, 0.681]	0.689 [0.632, 0.750]	0.848 [0.812, 0.881]

B SCORING AND STATISTICAL METHODS

Answer-label scores. The diagnostic proxy normalizes scores across answer letters. The seven proxy tokenizers represent both the plain and space-prefixed A/B/C/D labels with one token per spelling, without cross-letter token collisions. The historical estimator takes the higher score over these two spellings for each letter. This convention defines the proxy score and is not a claim of calibrated reader confidence. Model and tokenizer revisions appear in Appendix F.

The signed probability change Δp is defined in Equation 2. It is indexed by the exact definition text, its scope and its version. A mean log-score contrast, stored as `delta_mc` in the supplementary data, is a separate statistic and is not substituted for Δp.

Table 12: Bare error and proxy R AUROC by term class. Synthetic classes are fixed by the corpus design. Reddit glossary terms have a community entry, and control words are ordinary words from the same communities without one.

			DeepSeek Flash		GPT-6-astra	
Corpus	Term class	n	Error	R AUROC	Error	R AUROC
Synthetic	Redefined common word	156	0.442	0.803	0.365	0.777
Synthetic	Coinage	158	0.253	0.779	0.127	0.720
Synthetic	Usual meaning	142	0.148	0.921	0.042	0.846
Reddit	Glossary term	851	0.343	0.781	0.148	0.659
Reddit	Control word	540	0.235	0.842	0.143	0.815

Pairing and evaluation units. Prediction comparisons match record IDs and question stems, with context indices for the CUAD library. Reader records specify answer keys and option order. The main table reports question-level outcomes. Term-level Sonnet summaries have a different unit and remain in the supplementary inventory.

Definition-anchored CUAD task. On 671 matched questions, bare accuracy is 0.772 and accuracy with the drafted `glossC` is 0.903. The separate original-contract `glossG` arm reaches 0.982, and its correctness differs from `glossC` on 57 items. Error-ranking AUROC is 0.858 for R and 0.666 for Δp computed with the drafted gloss. These proxy probabilities do not measure the response to the contract text. This task is distinct from the 1,500-question usage validation.

Stratified validation. The five CUAD risk strata contain 2,529, 1,430, 1,085, 980, and 505 scored questions. A seeded shuffle with seed 42 selects 300 questions per stratum. The inverse inclusion weight is the stratum population divided by 300. Weighted error AUROC uses weighted positive-negative pairs. Within the 6,529-question frame, the weighted failure rate is 13.2%, R AUROC is 0.846, and Δp AUROC is 0.500. These weights do not cover occurrences excluded during question construction.

Uncertainty intervals. Primary intervals use 2,000 cluster-bootstrap replicates. Paired arms share each resample. CUAD clusters are contracts, Reddit clusters are communities, and synthetic and DeFi comparisons use term records. The intervals quantify variation across the observed clusters, not across arbitrary organizations, question generators, or repeated model calls. Exploratory subgroup and threshold results are interpreted separately from primary comparisons.

The Synthetic / GPT-6-astra and DeFi / DeepSeek V4.1 Flash rows in Table 2 round to the same R point estimate and interval at three decimal places. Their unrounded values differ: 0.8094577 with interval $[0.7627638, 0.8560215]$ and 0.8093046 with $[0.7625671, 0.8555733]$, respectively. The duplicate printed values result from rounding under the manuscript's original bootstrap procedure, rather than copying one row into the other. A separate bootstrap seed changes finite-replicate endpoints slightly and is not substituted into only one row.

Repair prediction. Table 13 evaluates prediction of a wrong-to-correct transition after a definition. Its final column reports a paired interval for the difference between the two scores on the separate bare-error target.

Algebraic coupling of error risk and definition response. Let p_0 and p_d be the proxy probabilities assigned to the keyed option without and with the supplied definition. By construction, $R = 1 - p_0$ and $\Delta p = p_d - p_0 = R + p_d - 1$. Because $0 \leq p_d \leq 1$, the two scores satisfy $R - 1 \leq \Delta p \leq R$. They are therefore neither algebraically nor statistically independent. The paired bootstrap interval in Table 13 compares their rankings on the same bare-error target. Pairing handles shared sampling variation but does not remove this structural dependence. It is not evidence that Δp isolates a distinct cause of error.

For the bare-error target, a Δp AUROC below 0.5 means that low definition response tends to accompany observed errors. In DeFi / GPT-6-astra, for example, mean R is 0.704 for bare errors and 0.494 for bare-correct items, but mean post-definition support p_d is 0.330 versus 0.651. The post-

Table 13: Prediction of observed repair. The last column concerns the separate bare-error target and gives a paired ranking difference. R and Δp are algebraically coupled (Appendix B).

Setting	n	R, repair AUC	Δp, repair AUC	Error AUC difference CI
Synthetic / DeepSeek V4 Flash	456	0.833	0.810	[-0.168, -0.046]
Synthetic / GPT-6-astra	456	0.797	0.716	[-0.259, -0.066]
Reddit / DeepSeek V4.1 Flash	1391	0.744	0.709	[-0.323, -0.244]
Reddit / GPT-6-astra	1391	0.725	0.621	[-0.364, -0.213]
DeFi / DeepSeek V4.1 Flash	379	0.769	0.686	[-0.445, -0.251]
DeFi / GPT-5.6-terra	379	0.792	0.556	[-0.502, -0.327]
DeFi / GPT-6-astra	379	0.754	0.590	[-0.510, -0.324]
CUAD library / GPT-5.6-terra	1500	0.760	0.595	[-0.395, -0.302]

Table 14: Why Δp can rank bare-reader errors below chance. Means are listed as bare-error / bare-correct items. The final column compares only error-correct pairs inside the same proxy-R decile, a descriptive approximate control for the common p_0 term. It is not a causal decomposition.

Setting	Δp error AUC	Mean p_d	Mean Δp	Within-R-decile AUC
Synthetic / DeepSeek V4 Flash	0.745	.804 / .924	.496 / .246	.485
Synthetic / GPT-6-astra	0.656	.751 / .920	.450 / .287	.431
Reddit / DeepSeek V4.1 Flash	0.520	.440 / .675	.163 / .131	.409
Reddit / GPT-6-astra	0.429	.405 / .639	.105 / .147	.359
DeFi / DeepSeek V4.1 Flash	0.458	.410 / .670	.116 / .130	.411
DeFi / GPT-5.6-terra	0.380	.356 / .668	.064 / .144	.307
DeFi / GPT-6-astra	0.342	.330 / .651	.034 / .144	.273
CUAD / GPT-5.6-terra	0.464	.407 / .714	.139 / .141	.398

definition difference more than offsets the risk difference, leaving mean Δp at 0.034 versus 0.144. The CUAD means nearly cancel, while pairwise ranks still give AUROC 0.464. A weak or negative association with bare error does not contradict a positive association with repair conditional on bare error, because those are different prediction targets. The within-decile values remain descriptive: they do not control item difficulty beyond R, question construction, or source-key quality. The more specific target restricts evaluation to a reader's bare failures and asks which of those errors the same definition fixes. Table 4 gives direct and within-term sibling estimates for the recorded intervention text. Conditional results cannot by themselves guide pre-answer selection, because bare failure is then unknown.

Question-independent meaning difference. We separately measure how unlike a local definition is from a model's prior gloss of the term. The uncontextualized distance is $1 - \cos(e(d_{\text{source}}), e(g_{\text{prior}}))$, where e is the reported BAAI/bge-base-en-v1.5 sentence encoder (Xiao et al., 2024). A second distance uses the model's domain-conditioned prior gloss. Model self-report scores ask whether the term has a special use in the corpus, and model-assigned relation labels include none, narrower, and shifted. These measurements do not use a diagnostic question. For contrast, we also show a term-mean comparison of an automatically drafted local gloss against an elicited prior gloss, together with term-mean Δp and R. Those three scores do use diagnostic questions. Distances are unavailable when the model returned an unknown prior gloss.

The strong synthetic separation is a positive control for a deliberately designed semantic contrast, not an estimate of real-corpus misreading. To compare a source-definition distance with reader behavior, we use only the archived reader arm that supplied that same source definition, `glossG`. A term's repair rate is the fraction of its historical questions answered wrongly without a definition and correctly with `glossG`. This term-level outcome uses a median of four questions per term. The automatically drafted `glossC` arm would compare a different intervention text and is excluded from this alignment.

The direct version-aligned test of the proposed term-level link uses the Reddit proxy's archived scores with the exact community entry. On 209 terms with 418 A questions, no-context distance has Spearman correlation 0.188 with the term's mean source-definition Δp, with a term-bootstrap interval $[0.047, 0.323]$. This is a weak positive association, not evidence that meaning difference is

Table 15: Distinguishing designed redefinitions from usual-meaning terms in the synthetic vocabulary, after excluding one source-definition-conflicted term. Each score uses all available terms for that score. Coverage differs and point estimates should not be ranked as if measured on one common sample. The class label was fixed by construction.

Term score	Uses questions?	Terms (shifted / consistent)	Class AUROC
No-context definition distance	No	66 (38 / 28)	0.946
Domain-prior definition distance	No	157 (118 / 39)	0.977
Model self-report	No	79 (40 / 39)	0.984
Drafted-local/prior answer contrast	Yes	78 (40 / 38)	0.917
Mean Δp	Yes	78 (40 / 38)	0.783
Mean R	Yes	78 (40 / 38)	0.817

Table 16: Spearman association of question-independent term scores with source-definition repair rate. Parentheses give available terms. Distance confidence intervals resample term records. There is no aligned `glossG` reader arm for this 204-term CUAD package. Only 54 terms have an uncontextualized prior gloss.

Corpus	No-context distance [95% interval]	Domain-prior distance	Self-report
Synthetic	0.452 $[0.260, 0.602]$ (65)	0.300 (186)	0.465 (118)
Reddit	0.181 $[0.046, 0.303]$ (211)	0.095 (217)	0.166 (217)
DeFi	-0.087 $[-0.361, 0.186]$ (57)	0.101 (64)	-0.237 (64)

absent. It cannot identify which future question falls on a decision boundary. Both this proxy test and the repair test still use historical model-drafted questions, and the separate source audit finds substantial answerability uncertainty for Reddit and DeFi.

Model-assigned relation labels also give weak and inconsistent repair separation on the source-definition arm. Among Reddit glossary terms, 186 labeled narrower or shifted have mean repair rate 0.292, versus 0.266 for 31 labeled none. The term-level repair AUROC is 0.531. In DeFi, the corresponding rates are 0.227 versus 0.333 over 32 terms per group, with AUROC 0.411. These are model-assigned lexical relations, not independent annotations of the three situations in Table 1. A single text distance cannot by itself tell whether a use involves meaning conflict, an unspecified local scope, or missing information elsewhere. The synthetic reader-aligned text pairs and embedding script are retained, and the other source/prior texts remain in the archived weights. The exact BGE revision and embedding vectors are not, so exact bitwise reproduction of the cosine values is not guaranteed. The reported associations and source-aligned reader outcomes are reproducible from the archived scores.

Sampling consistency. The archived synthetic uncertainty experiment contains ten sampled responses per question, for 7,630 answer letters, and a separate greedy response. After removing the two source-conflicted questions, the analysis uses 761 questions and 7,610 of the archived letters. Among 229 greedy errors, 154 have ten identical sampled answers. Of these, 93 are identical and wrong, while 61 are identical and correct. The majority sampled answer is wrong on 142 of the 229 items. The corrected common proxy comparison uses 456 questions. Entropy is computed over option votes, and elicited confidence is treated as a response measurement rather than a calibrated probability.

C Definition interventions

Lexical shortcuts in generated questions. The diagnostic questions and local glosses can share words even when the question does not quote the gloss verbatim. A simple warning rule removes the target term and common English words, then chooses the option with the unique largest fraction of content words overlapping the exact drafted `glossC`. The synthetic `raw_outputs` glosses match all 197 recovered reader term prompts. A different same-named input file does not. Table 18 reports how often this rule selects the stored key. A shuffled-gloss baseline permutes glosses among terms

Table 17: DeepSeek Flash readers on matched questions under four conditions: V4 for synthetic and V4.1 for Reddit and DeFi. Cell values are accuracy, followed by repairs/harms relative to bare. `glossG` is the source definition, `glossC` the model-drafted local gloss, and `glossM` a domain-conditioned prior gloss elicited without the source examples. The synthetic row excludes two source-conflicted questions. Reddit and DeFi drafters saw the source definition.

Corpus	n	Bare	Source `glossG`	Drafted `glossC`	Domain-prior `glossM`
Synthetic	761	.699	.966 (209/6)	.949 (196/6)	.708 (86/79)
Reddit glossary	851	.657	.932 (243/9)	.790 (171/58)	.616 (82/117)
DeFi source-entry	250	.672	.940 (70/3)	.760 (42/20)	.740 (40/23)

within the corpus. The reader's own question is retained in the no-cue subset, but a missing lexical flag is not a semantic proof that no answer leakage remains.

Table 18: Lexical-cue screen using the exact `glossC` text supplied to the archived DeepSeek Flash readers (V4 synthetic. V4.1 Reddit and DeFi). Synthetic values use the reader-aligned `raw_outputs/v1_weights.jsonl`, rather than a different same-named input copy. Error rates are bare $\rightarrow$ drafted-gloss.

Corpus	n	Cue	Shuffle	No-cue n	No-cue errors	No-cue AUC: R / repair-Δp
Synthetic	456	.686	.067	143	.308 $\rightarrow$ .091	.879 / .950
Reddit	1,391	.185	.032	1,133	.296 $\rightarrow$ .210	.810 / .844
DeFi	379	.274	.086	275	.284 $\rightarrow$ .236	.803 / .791

The high synthetic cue rate means the full-sample 0.285 $\rightarrow$ 0.029 error reduction is partly exposed to an option-matching shortcut. The no-cue subset still shows a substantial change, but its definition arm may contain unmeasured semantic or structural cues. Reddit and DeFi have lower lexical-cue rates, yet their question writers saw source definitions and their answer-key audits have additional limitations.

Definition arms without lexical cues. For the four-arm synthetic comparison, the same warning rule is applied separately to the source definition, the drafted gloss, and the model's own gloss. The rule flags 45.2%, 63.1%, and 29.3% of the 761 questions for the three texts. On the 180 questions flagged by none of them, DeepSeek V4 Flash accuracy is 0.617 without a definition, 0.911 with the source definition (56 repairs, 3 harms), 0.883 with the drafted gloss (52 repairs, 4 harms), and 0.544 with its own gloss (15 repairs, 28 harms).

Retrieval configurations. CUAD compares BM25, a dense BGE retriever with reranking, and hybrid BGE-M3 retrieval (Robertson & Zaragoza, 2009; Xiao et al., 2024; Chen et al., 2024). The configurations use eight retrieved passages. Reddit uses six passages. The CUAD full-glossary reference is capped at 30 entries. Risk selection uses a term maximum within each contract and a term mean within each Reddit community.

Overlap between scoring and evaluation questions. In both hybrid workloads, the term aggregate can include the evaluated question's own diagnostic score. For CUAD, the selected set remains unchanged after omitting that score on 990 of 1,200 questions. 17 changed sets become empty and have an observed retrieval-only response. The remaining 193 changed sets have no matching reader arm. Their unobserved binary outcomes yield a partial-identification interval $[0.733, 0.893]$ for leave-one-question-out accuracy, compared with 0.821 for retrieval. For Reddit, 1,295 of 1,391 gated decisions are unchanged, 33 reduce to retrieval-only, and 63 lack the necessary reader response. The corresponding interval is $[0.874, 0.919]$ over all questions, versus 0.866 for retrieval. On the 851 glossary questions it is $[0.859, 0.933]$, versus 0.839 for retrieval. These intervals put 0 and 1 on unobserved outcomes. Substituting archived retrieval or in-sample gated answers for them gives narrower scenarios, not valid bounds. They are finite-workload identification intervals, without clustered sampling uncertainty. A question-level exclusion is not a held-out term or document split. Random and frequency policies at the same budget, and a Reddit all-matched arm, are not observed in this workload.

Table 19: In-sample hybrid-retrieval results on constructed questions. The risk aggregate includes the evaluated diagnostic item. The all-matched arm is unavailable for Reddit. The CUAD full glossary is capped at 30 entries. These arms are not matched-budget random or frequency policies.

Workload	n	Retrieval	Selected notes	All matched	Full glossary
CUAD	1,200	0.821	0.871	0.882	0.876
Reddit / all	1,391	0.866	0.925	n/a	0.907
Reddit / glossary	851	0.839	0.941	n/a	0.907
Reddit / control	540	0.907	0.898	n/a	0.907

The additional-passage arm is approximately cost-matched because it adds whole passages. The all-matched arm adds definitions for detected terms without risk gating. Definition lengths in Table 20 exclude retrieval-only context and use the same regex counter within each workload.

Table 20: Full RAG accounting: selected-definition policy relative to retrieval only. Repairs and harms are paired counts. Added definition units are inferred by subtracting identical retrieval contexts. They include glossary formatting.

Corpus / tier / subset	n	Gain, pp [95% CI]	Repair/harm	Selected/full units
CUAD / BM25 / all	1200	4.17 [2.24, 6.12]	94 / 44	135.8 / 529.2
CUAD / dense / all	1200	5.33 [3.49, 7.35]	102 / 38	79.1 / 529.2
CUAD / hybrid / all	1200	5.00 [3.34, 6.73]	93 / 33	126.6 / 529.2
Reddit / BM25 / all	1391	6.69 [4.45, 9.04]	112 / 19	39.4 / 2765.9
Reddit / BM25 / glossary	851	10.34 [7.01, 13.83]	97 / 9	46.5 / 3285.2
Reddit / BM25 / control	540	0.93 [-0.89, 2.78]	15 / 10	28.3 / 1947.5
Reddit / hybrid / all	1391	5.90 [3.85, 7.77]	106 / 24	38.5 / 2765.9
Reddit / hybrid / glossary	851	10.22 [7.61, 13.30]	92 / 5	46.9 / 3285.2
Reddit / hybrid / control	540	-0.93 [-2.86, 1.12]	14 / 19	25.3 / 1947.5

Inline selection. The 965 synthetic probes have 85 repairs and four harms. The 733 CUAD probes have 29 repairs and 16 harms. Full-injection net gains are 8.39 points with interval $[5.73, 11.46]$ and 1.77 points with interval $[-0.44, 3.91]$. An offline selection curve sums paired effects on selected probes and divides by the entire workload size. The random-allocation expectation is the selected fraction times the full-injection effect.

Other diagnostic comparisons. Self-report separates designed synthetic term classes more strongly than it predicts errors on real corpora. Frequency, word-list rarity, and contextual embedding shift have weaker reported associations with reader failures. Masked-span and span-likelihood estimators combine completion difficulty with local-meaning dependence. Sentence surprisal has little reported association with this risk target.

A sampled Haiku proxy has lower reported discrimination than the local model on its covered subset. Jev, released by TypeSafe AI (Almeida, 2026), provides native decision probabilities and confidence. Using negative native confidence to rank Jev's own bare errors gives AUROC 0.643 on 761 synthetic questions (the same two source-conflict items are excluded as elsewhere), 0.660 on 1,391 Reddit questions, and 0.636 on 671 CUAD questions. When a model-drafted local gloss is prepended, the same confidence ranks Jev's errors with AUROC 0.815, 0.790, and 0.868 on the three sets, and Jev's own error falls from 0.418 to 0.046 on the synthetic set, 0.423 to 0.206, and 0.323 to 0.113. These are own-error measurements for a different decision model, not direct cross-reader comparisons with R. Model-size, option-order, and repeated-question-set experiments are listed in the supplementary inventory. Limited overlap between readers' errors does not establish strict nesting or an asymptotic error floor.

Explicit Inverse Scaling redefinition items supply the local meaning in the prompt. Their compliance results concern following an available definition (McKenzie et al., 2023; Stringli et al., 2025).

Decoding, scope, and carriers. The white-box decoding experiment uses Qwen2.5-14B-Instruct as reader on 1,200 CUAD questions under hybrid retrieval and compares greedy decoding, CAD, and AdaCAD (Shi et al., 2024; Wang et al., 2025a). With retrieved passages only, CAD accuracy is 0.719 against 0.747 for greedy decoding, a change of -2.8 points with contract-cluster interval $[-4.4, -1.0]$. With two same-contract usage examples added, the change is -2.3 points $[-4.0, -0.7]$. With risk-selected definitions added, greedy and CAD accuracies are 0.818 and 0.816, a change of -0.2 points $[-1.7, +1.3]$. This outcome is expected from the construction of contrastive decoding, which reweights information already in the input. It confirms, for one white-box reader, that faithfulness-oriented decoding does not supply a missing definition. These results concern the tested configurations.

Plain, scoped, and negatively framed glossaries produce different rates of abstention on ordinary unlisted terms and over-application to usual-meaning questions. Scope wording reduces some of these effects. The carrier study compares full-context text, retrieval-scoped text, LoRA, and a GRACE-style editor (Hu et al., 2022; Hartvigsen et al., 2023). Outcomes of individual configurations do not establish a general limitation of either method family.

D PUBLIC HUMAN-WRITTEN TASKS

Table 21: Public-task interventions. Metric changes are in raw F1/accuracy units, with paired cluster-bootstrap intervals. Reddit labels refer to post-term items.

Task	n	Arm	Score	Difference [95% CI]
CUAD F1	1661	rag_full	0.5270	+0.0029 [-0.0037, +0.0100]
CUAD F1	1661	rag_gated	0.5249	+0.0008 [-0.0051, +0.0073]
ContractNLI acc.	3060	note_gated	0.8173	+0.0059 [-0.0020, +0.0147]
ContractNLI acc.	3060	note_all	0.8222	+0.0108 [+0.0026, +0.0193]
Reddit / GPT	178	note_gated	0.5169	+0.0169 [-0.0382, +0.0566]
Reddit / GPT	178	note_all	0.5281	+0.0281 [-0.0715, +0.0909]
Reddit / DS	178	note_gated	0.6124	+0.0056 [-0.0335, +0.0473]
Reddit / DS	178	note_all	0.5955	-0.0112 [-0.0592, +0.0373]

The CUAD extraction comparison uses 1,661 matched questions, and adding definitions changes token F1 by at most 0.3 points. ContractNLI (Koreeda & Manning, 2021) uses 3,060 hypotheses with document clustering. Adding all matched notes raises accuracy by 1.1 points with an interval above zero, and gated notes by 0.6 points. Both tasks ask what a clause states rather than whether a case falls inside a defined term, so few items depend on a definition's boundary. The Reddit human-post comparison contains 178 post-term items from 176 posts. Source posts and responses are human-authored, while the evaluated multiple-choice questions are pipeline-produced. The TechQA check contains 105 drafted questions on 20 screened terms rather than the complete human-authored support task.

E DEFINITION SELECTION AND SOURCE-TEXT SENSITIVITY

CUAD development allocation. The 63-question batch is drafted separately from its A diagnostics but remains a small, constructed, forced-choice task. Full-definition provision yields 61 versus 51 correct answers for GPT-5.6-terra and 59 versus 51 for DeepSeek V4.1 Flash. At the 50% fixed question budget, mean Δp, matched-first, mean R, and frequency yield 55, 56, 57, and 59 correct GPT-5.6-terra answers. The corresponding DeepSeek V4.1 Flash counts are 56, 55, 55, and 56. No single Δp policy dominates these simple comparators across readers and budget points. This batch is not pooled with the Reddit library experiment.

Shared-library split and definition versions. The Reddit calculation uses 215 glossary terms in 35 communities. Questions of a term are split by odd and even index into diagnostic A and evaluation B, so B is a sibling question rather than an independent future request. Each half usually has one or two questions per term. Frequency comes from the historical corpus. The library-policy comparison uses the exact community entry for A scoring, word cost, and B reader intervention.

A separate drafted-gloss arm uses a different text for both scoring and intervention. Every drafted gloss differs from the corresponding community entry.

The proxy scores 429 A questions and the reader outcomes cover 422 B questions. The table below gives AUROC for predicting a wrong-to-correct transition conditional on a bare error. The A direct row scores the same question. The A-to-B row uses the A-term mean to score the other half. Intervals resample communities.

Table 22: Reddit conditional repair with the version of the definition held fixed between proxy scoring and reader intervention. These historical questions were model drafted and are not independent future requests.

Definition text	A direct: bare failures / repairs, AUROC [95% CI]	A to B: bare failures / repairs, AUROC [95% CI]
Model-drafted gloss	144 / 85, 0.833 [0.773, 0.882]	147 / 85, 0.811 [0.732, 0.893]
Community source entry	144 / 116, 0.665 [0.539, 0.788]	147 / 126, 0.600 [0.475, 0.717]

The community source entry repairs more bare failures and causes fewer correct-to-wrong changes: the latter are four on B, versus 25 for the drafted gloss. Its near-ceiling B repair rate leaves only 21 unrepaired bare errors. This small negative class increases the uncertainty of the ranking estimate, but cannot by itself explain the 0.600 point value. The two definition versions have different AUROCs because Δp applies only to the text it scores. The B questions were drafted with the community entry and their keys have not all been independently adjudicated, so the comparison is not a randomized definition-quality experiment.

Configured reader and answer quality. The source-entry 25% word-budget comparison selects 111 definitions by frequency per cost and 89 by frequency times Δp per cost. Only 60 terms differ. GPT-5.6-terra supplies bare and source-entry responses for 118 B questions on these terms, producing 236 model responses with recorded usage and response IDs. The frequency-only terms and Δp-only terms replace 549 and 550 definition words, respectively. They yield 26 versus 11 wrong-to-correct changes, with one versus zero harms. Under the historical frequency weights, the Δp policy has 4.53 fewer correct outcomes per 100 eligible occurrences. A community-cluster interval conditional on the fixed two libraries is $[-9.45, -0.28]$ points. It does not account for uncertainty in source-validity decisions or independent future use.

An outcome-blind DeepSeek V4.1 Flash source review accepts 83 of the 118 B questions under a fixed rule requiring one supported answer and no answer leakage. Fifteen differing terms have no accepted question. Bounding each missing term's mean repair effect by $[-1, 1]$ gives a total difference interval of $[-9.21, +6.14]$ points. Model acceptance is not expert gold, and a fixed-hash spot review finds some remaining underspecified questions. The available evidence does not establish which library is better for future requests.

F MODELS AND REPRODUCIBILITY

The GPT reader identifiers are `gpt-5.6-terra` and `gpt-6-astra`. The experiment operator confirms that synthetic DeepSeek reader calls made before September 10, 2026 used V4 Flash, while later Reddit and DeFi reader calls used V4.1 Flash. DeepSeek documents `deepseek-flash` as the current V4.1 API ID and says that the legacy `deepseek-v4-flash` ID routes to V4.1 after the September 10 release (DeepSeek, 2026). The later source-audit model is also V4.1. We do not assign a knowledge cutoff to either DeepSeek version because no official cutoff was established for these experiments. CUAD's 1,500-question validation uses `gpt-5.6-terra`. The additional synthetic, Reddit, and DeFi readers use `gpt-6-astra`. Historical proxy revisions are listed below.

Reader diagnostic calls use greedy-style decoding, with temperature zero where supported. The uncertainty protocol uses temperature 1.0 and ten samples per item. The proxy scores answer labels from local model logits. RAG costs are reported as context-length units rather than API billing tokens. Sonnet comparisons based on term means are reported at term level. After the source-

Table 23: Proxy model revisions. Full commit identifiers are supplied in the accompanying model records.

Model	Revision prefix
Qwen2.5-0.5B	`060db6499f32`
Qwen2.5-1.5B	`8faed761d45a`
Qwen2.5-3B	`3aab1f1954e9`
Qwen2.5-7B	`d14972939875`
Qwen2.5-14B	`97e1e76335b7`
Qwen2.5-14B-Instruct	`cf98f3b3bbb4`
Llama-3.1-8B	`d04e592bb4f6`

conflict exclusion, response-level coverage of the GPT-6 synthetic usual-gloss condition is 440 of 456 items, while bare and local-definition responses cover all 456.

The development implementation records request settings, source and question hashes, returned model fields, token usage, response status, and per-policy outcomes. Its code and result export exclude credentials. Construction and maintenance of the diagnostic index are separate costs from supplying notes to a reader. Any deployment-cost analysis must account for reuse over subsequent requests.

G Exploratory analyses

The analyses below did not yield a conclusive finding and are reported for completeness.

Clause alignment. The CUAD alignment analysis covers 5,659 scored occurrences, including 644 inside expert-labeled clause spans and 5,015 outside. The share with proxy probability response above 0.3 is 0.1134 inside and 0.1163 outside. Clause categories identify legal topics, not a scale of monetary or decision severity. Within-category comparisons are exploratory and categories may overlap.

Unrepaired errors. Of the 157 CUAD errors that remain wrong with a definition, automated annotations classify 53 as prior-winning interpretations, 39 as question defects, 35 as relation gaps, 15 as insufficient definitions, and one as a parameter issue. Fourteen are unclassified. These are model-generated descriptions, not established causal explanations. The present diagnostic and selection tests do not validate these causal labels.

H Supplementary experiment inventory

A machine-readable inventory of every experiment, with its evaluation unit, source hashes, and stored results, is provided in the supplementary materials (`experiment_inventory.tsv`).