The Double Edge of AI Annotation in Cultural Sciences

Edgar Dubourg[1,2,*] & Valentin Thouzeau[1]

[1]Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France

[2]Institut Curie, PSL Research University, 26 rue d'Ulm, 75248, Paris Cedex 05, France

* Corresponding author: edgar.dubourg@gmail.com

Abstract

Artificial intelligence is transforming the cultural sciences, letting researchers code text, images, and sound at a scale that was previously out of reach. But it raises a question current practice cannot answer: how much annotation error can a study tolerate before its conclusions stop holding? The usual reflex, demanding near-perfect agreement with human coders, sets a threshold that is disconnected from the analysis. We reframe the question in terms of inference: an AI annotation is good enough when its errors do not change the conclusion you would have reached with ground-truth annotation. Using Monte Carlo simulations across thousands of realistic designs, we show that the required annotation quality is a property of the study, not a universal threshold, and that it falls as the sample grows: with enough data, even a noisy annotator recovers a true effect, and random hallucination is no more damaging than ordinary noise. Systematic error is different: when mistakes correlate with the hypothesis instead of being random, they can create a confirmation bias, and a larger sample makes things worse rather than better. The two findings form a double edge: more data lowers the accuracy you need to find real effects, but raises the risk of validating false ones. We give cultural scientists a concrete way to compute, for their own design, both the size of the human audit and the correlation with AI annotations needed to keep the planned test valid.

Significance statement

Researchers across the cultural and social sciences now use artificial intelligence to label huge collections of text, images, and sound, measuring things like sentiment, theme, or tone that once required human annotators. But how accurate must these automated labels be before a study's conclusions can be trusted? Common practice demands near-perfect agreement with human coders. We show that the accuracy a study needs is a property of that study, not a universal cutoff, and that it relaxes as datasets grow. Yet the same abundance of data can turn small errors into false discoveries. We give researchers a way to calculate, for their own project, how good their AI annotations must be to avoid both false positives and false negatives.

1. Introduction

For most of their history, the cultural sciences (anthropology, history, literary studies, sociology, psychology, religious studies, etc.) were rate-limited by the human eye. To know what a corpus of texts, images, songs, or films was about, someone had to read, watch, or listen to all of it and write down what they saw. Coding a few hundred items was a very long and demanding task, and coding tens of thousands was rarely feasible at all.

This constraint had already begun to loosen before the current wave of artificial intelligence (AI). Over the past two decades, computational and quantitative methods opened a first window onto culture at scale. In literary studies, distant reading proposed to understand literature not by reading individual works closely but by analysing thousands of them at once (Moretti, 2000; Jockers, 2013), an approach that has since traced large-scale changes in literary form and content across centuries (Sobchuk, 2018; Underwood, 2019). In the cultural sciences, the digitization of books and records made it possible to follow the frequency of words and ideas through time, revealing large-scale patterns (Michel et al., 2011; Morin & Acerbi, 2017; Baumard et al., 2022). These approaches established that culture can be studied quantitatively and at scale. But they were largely restricted to what a computer could count on its own, while anything that required interpreting context or meaning still needed a human coder.

AI annotation has lifted that limit (Grossmann et al., 2023; Bail, 2024; Dubourg et al., 2024). An LLM can read a thousand folktales, a century of obituaries, or ten million tweets and assign each one a score on whatever dimension a researcher cares about: how violent it is, how much it appeals to fairness, how romantic, how religious, how funny (e.g., Dubourg et al., 2025). On many such tasks these models already match or exceed trained human coders (Gilardi et al., 2023; Ziems et al., 2023; Rathje et al., 2024). Questions that were once unaskable, because no one could annotate the data, are now within reach (Nakawake & Sato, 2019; Karjus, 2024; Dubourg et al., 2024), and this opens up a large number of new research questions.

But there is a catch, and it is not the one most people worry about. The problem is not that LLMs make mistakes, since every measurement instrument does. It is that we have no principled way to decide how many mistakes are acceptable. Faced with this, the field has instinctively borrowed a habit from inter-rater reliability: validate the model against a human gold standard and demand high agreement (e.g., Cohen's $\kappa \geq .80$; intraclass correlation $\geq .75$; Krippendorff's α above some convention; see Cohen, 1960; Krippendorff, 2018; Pangakis et al., 2023).

This paper argues that these universal cutoffs answer the wrong question. They treat "good enough" as a property of the AI annotation, fixed once and for all. But it is not: whether an annotator is good enough depends on the study you are going to run with it. A noisy annotation can be perfectly adequate for one analysis and dangerously inadequate for another, on the very same data. And an AI annotator can clear every conventional reliability threshold and still produce a result that is both statistically significant and false (Baumann et al., 2025).

Suppose you are a folklorist with a hypothesis: over the centuries, traditional folktales have become more romantic, with later tales devoting more attention to romantic love than earlier ones. You have collected 1,000 tales with known dates. To test the idea you need, for each tale, a

score for how much romantic love it contains, say on a 0-to-6 scale. Reading and scoring 1,000 tales by hand would be too long and effortful, so you ask an LLM to do it, and manually code a random sample of 100 tales, to check the model against your own judgement (Pangakis et al., 2023). This is a possible study: 1,000 tales annotated by the model (which we call N), 100 audited by hand (which we call n), one hypothesis (recency predicts romance) and one test (a regression of romance on date). The question we ask is simple: given this study, how good does the LLM's romance score have to be for the results and inference to be valid?

2. General method

Both of our studies use a generative model. Imagine first a perfect annotation, which we call the 'ground truth'. For each item i (each folktale) there is a predictor X_i (the tale's date, standardized) and a true latent quantity L_i (how romantic the tale really is). Our hypothesis is that the two are linked,

$$L_i = \beta \cdot X_i + \varepsilon_i,$$

where β is the effect size we care about (whether recency predicts romance, and how strongly) and ε is ordinary between-item variation. The ground-truth romance score is L binned onto the 0–6 scale the rater uses.

Now replace this ground truth with an LLM annotation. The proxy does not report L, but instead L corrupted by an AI annotation error. We study two kinds of error.

The first is the false negative: the proxy's score is the truth plus noise, either continuous jitter (the model is imprecise) or, with some probability, a random substitution (the model hallucinates a score unrelated to the text). A real effect is still present in the truth, but the noise dilutes it until the test can no longer detect it, and the researcher concludes "no effect" when there is one. This is the subject of Study 1.

The second is the false positive. If the model's errors are not random but systematic, leaning even slightly in the direction of the hypothesis (perhaps the model has the same prior as the researcher and reads a little more romance into tales it knows are recent), the error does not blur the signal but *generates* one, and the researcher concludes there is an effect when there is none. This is the subject of Study 2.

Crucially, the question is not whether the annotation is valid in the abstract, but how much annotation error a study can tolerate while its planned analysis still detects a true effect and still rejects a false one. We answer it by simulation. We build a world in which the truth is known, run the study, and turn the annotation error up until the analysis can no longer recover the answer it was designed to find. The point at which it breaks tells you how good your annotation has to be, expressed as the correlation you would need to observe with your manually annotated items. We use Spearman's ρ because the annotation scale we create is ordinal (the spacing between its values is arbitrary, so only the ranking of items carries information, and ρ is the measure that depends on that ranking alone).

To summarize, in each study, we generate the data so that we already know the right answer, precisely so we can check whether the annotation recovers it. In Study 1 we plant a real effect (romance increasing with X by a known β) and test whether error makes us miss it (i.e., a false negative). In Study 2 we plant no effect and test whether error makes us see one (i.e., a false positive).

3. Study 1: Avoiding false negatives (random error)

Study 1 asks how far annotation error can go before a real effect disappears. We explore this across many designs: the true effect ranges from weak to strong ($\beta \in \{.05, .1, .2, .5\}$), the corpus from 100 to 100,000 LLM-annotated items, and the human audit from a handful of hand-coded pairs up to 1,000. For each combination, we increase the annotation error until the study fails, repeating each cell 12,000 times (Monte Carlo) so the numbers are stable. We use conventional thresholds (i.e., significance at $\alpha = .05$ and adequate statistical power at 80%).

To exemplify, take a single study design first. Once its effect size β, its sample size N, and its audit size n are fixed, we can ask how much random annotation error it tolerates. We inject increasing error into the annotator, track the study's statistical power (its probability of detecting the real effect), and find the error level at which power falls through the conventional 80% floor. We then read off the annotation quality at that point, the Spearman correlation the model's scores would show against the human scores in an audit of n items. That correlation is the quality this particular design requires: any AI annotation above it recovers the real effect at least 80% of the time, and any annotator below it does not. In other words, for the study we took as example, any LLM whose romance scores correlate at least about 0.59 with human judgement is good enough (**Figure 1**). Scored on a conventional reliability index, that same AI annotation falls short: Krippendorff's alpha, for instance, comes out at about 0.43 on ordinal scores, well below the .80 usually read as adequate. We would therefore discard an annotation that is entirely adequate for the test at hand.

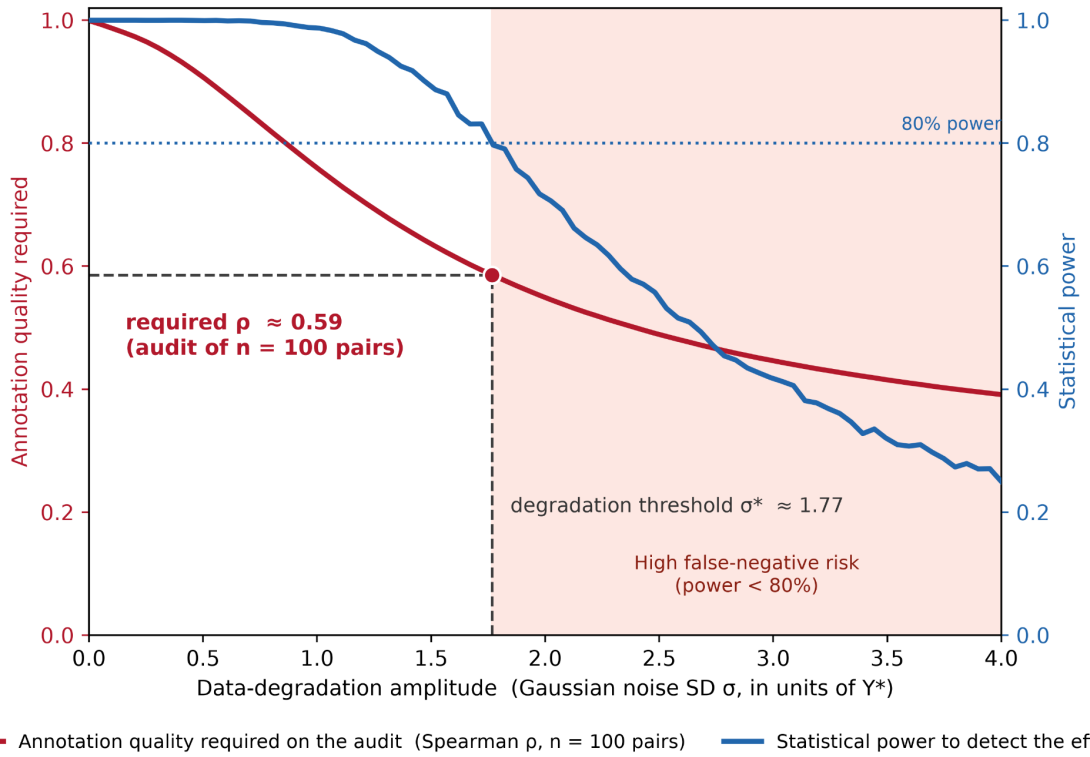

Figure 1. Required annotation quality for a single design, here the folktale study (continuous random error, β = 0.20, N = 1,000, n = 100). The blue curve (right axis) is statistical power as annotation error increases; it crosses the 80% floor at the noise level $\sigma^* \approx 1.77$. The red curve (left axis) is the Spearman correlation the annotator would show in the audit at each error level. Reading down from the power crossing to the red curve gives the quality this design requires, $\varrho \approx 0.59$.

This example is about one study; we run the same procedure for every design specification. **Figure 2** lays the required-quality number out across the whole grid. Each cell is a study design: its sample size N is on the vertical axis (log scale), its audit size n is on the horizontal axis (log scale), the four columns are the four effect sizes β, and the two rows are the two kinds of random error (top: ordinary noise; bottom: hallucination). Two regions of the map are greyed out, because in them no AI annotation, however accurate, can safely test the hypothesis. The first, "Underpowered by design" (grey, hatched), includes studies too small relative to the effect they look for (therefore, in them, even a perfect annotation ($\varrho = 1$) cannot reach 80% power). The second is the diagonal where the audit n is as large as the corpus N: hand-coding as many tales as the model annotates makes the model redundant, so the LLM is only useful when N is larger than n. Outside these two zones, the colour of each cell is the target quality for that study.

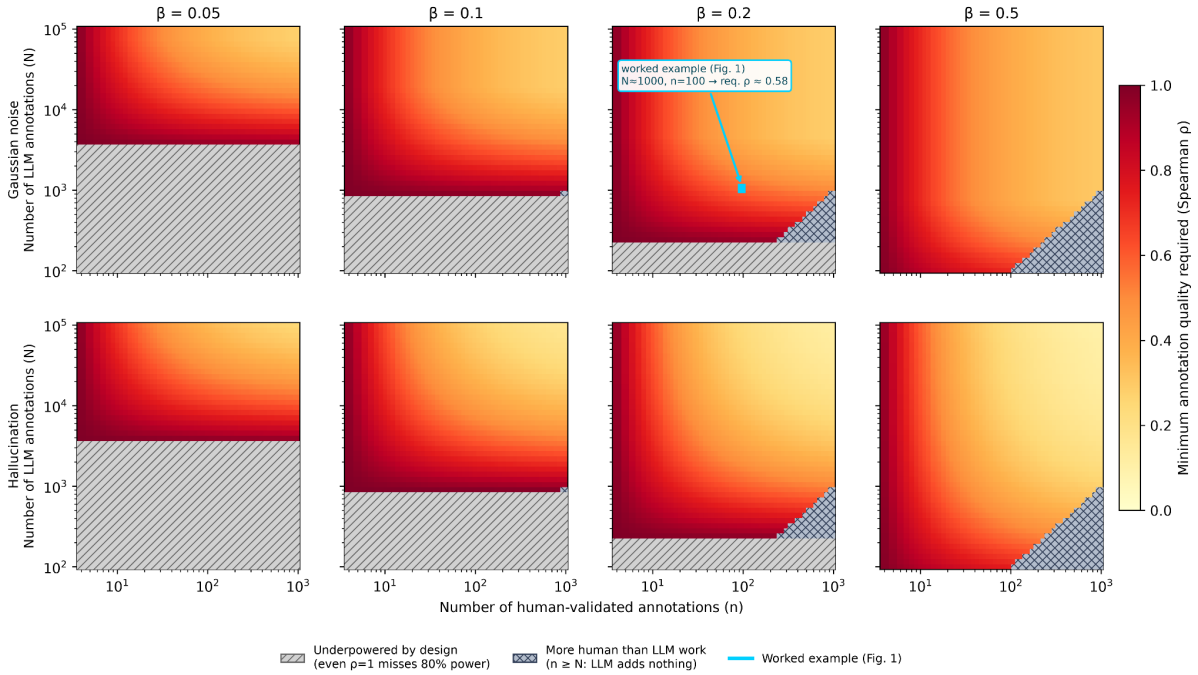

Figure 2. Required annotation quality across all designs. Rows: the kind of random error (top, continuous noise; bottom, hallucination). Columns: the effect size β. Within each panel, the corpus size N runs up the vertical axis (log) and the audit size n along the horizontal axis (log). Cell colour is the Spearman ϱ the design requires, from pale yellow (lenient) to deep red (demanding). Grey hatched cells are underpowered by design (even ϱ = 1 cannot reach 80% power); blue cross-hatched cells have n ≥ N (the LLM adds nothing). The cyan circle marks the folktale study of Figure 1 (ϱ ≈ 0.59).

The result is intuitive: holding the effect size fixed, a larger sample lowers the annotation quality required to detect it. This follows the logic by which large samples detect small effects. Random error shrinks an effect estimate toward zero (Spearman, 1904; Carroll et al., 2006), but a large enough sample can still resolve the shrunken effect with confidence. With 100,000 tales one could tolerate a far noisier annotation than with 1,000. This is good news: if you can annotate more, you can annotate worse.

A second result concerns hallucination. One might expect that a model occasionally inventing a score is a categorically worse problem than ordinary imprecision. The bottom row of Figure 2 shows it is not. Random substitution behaves almost exactly like continuous noise as the underpowered zone is virtually identical. If anything, hallucination is slightly less demanding, with the required ϱ dropping to about .045 for the largest effect and sample. As long as the hallucinated scores are not correlated to the independent variable X (e.g., the date, see study 2), they are just another source of random dilution. They cost power like noise, and nothing more (Pangakis et al., 2023; Rathje et al., 2024). Random error of any kind is a false-negative problem and never a false-positive one.

4. Study 2: Avoiding false positives (confirmation bias)

Study 1 assumed the annotator's errors were random with respect to the predictor. When that assumption fails, the danger changes (Boelaert et al., 2025). Return to the folktales, but suppose this time that our hypothesis is false: recent tales are no more romantic than older ones. There is no real effect to find. Now, suppose the AI does not merely score romance imprecisely but

scores it with a tilt that aligns with the hypothesis: it reads slightly more romance into tales it can tell are modern, because it hypothesizes (as the researcher has) that romance is more central in modern stories (Baumard et al., 2023). A tilt of this kind is not noise but a signal, pointing in the hypothesized direction. The regression will then return a slope, recent tales scoring higher on romance, but that slope reflects the AI annotation's bias rather than a property of the folktales. The result is a false positive, reported with a small p-value and full confidence, because the effect is real in the annotations even though it is absent in the world.

What makes this danger specific is that it comes down to one thing: whether the annotation error is correlated with the predictor X. If that correlation is zero, we are back in Study 1, where only power is at stake. If it is negative (in the folktales, the AI reads systematically less romance into recent tales than into older ones) it can only pull the estimate downward. In this case, the test becomes even more conservative: if a real effect does exist, this negative bias can potentially mask it, turning the problem back again into one of power; if the effect is still detected, though, the true effect must be even larger than the estimate. Validity is threatened only in the remaining case, when the error leans in the same direction as the effect we are looking for, creating a positive signal that was never there.

To put a number on this risk, we first need to measure the bias itself. We call it γ: the degree to which the annotation error tracks the predictor. Concretely, γ is the slope you obtain by regressing the annotation error (the model's score minus the human score in the audit) on X. When $\gamma = 0$ the error is random and we are back in Study 1; when γ is positive the error leans in the hypothesized direction, and it adds to the measured slope a spurious component proportional to γ (e.g., the model scores romance higher than human coders do *in recent folktales but not in older ones*). This is exactly the quantity an audit can estimate, from the hand-coded pairs, by regressing the model-minus-human difference on X. The question for any design is then how large γ can be before it produces a significant effect on its own.

We run the same kind of simulation as in Study 1 (Figure 3). The colour shows that threshold, γ_{fatal}: the smallest amount of hypothesis-aligned bias sufficient to push a typical study over the significance threshold, that is, the bias at which the false effect lands, on average, exactly on $p = .05$. Studies below γ_{fatal} are generally safe and studies above it generally report a spurious effect. Because γ_{fatal} depends only on the sample size N, the colour forms horizontal bands.

At the bottom (small N, green) studies are more robust: a sizeable bias, up to about $\gamma \approx 0.21$ at $N = 100$, is needed to create a fatal false positive. In other words, a small study is hard to fool. At the top (large N, deep red) studies are fragile: at $N = 100{,}000$ a bias of just $\gamma \approx 0.006$ (far below anything a validation would flag) is fatal. A large sample that tolerates random error will instead amplify systematic error (see Meng, 2018). With a large N, the confidence interval around the slope is narrow, and it is just as narrow when the slope is biased. A small bias is then enough to reach significance.

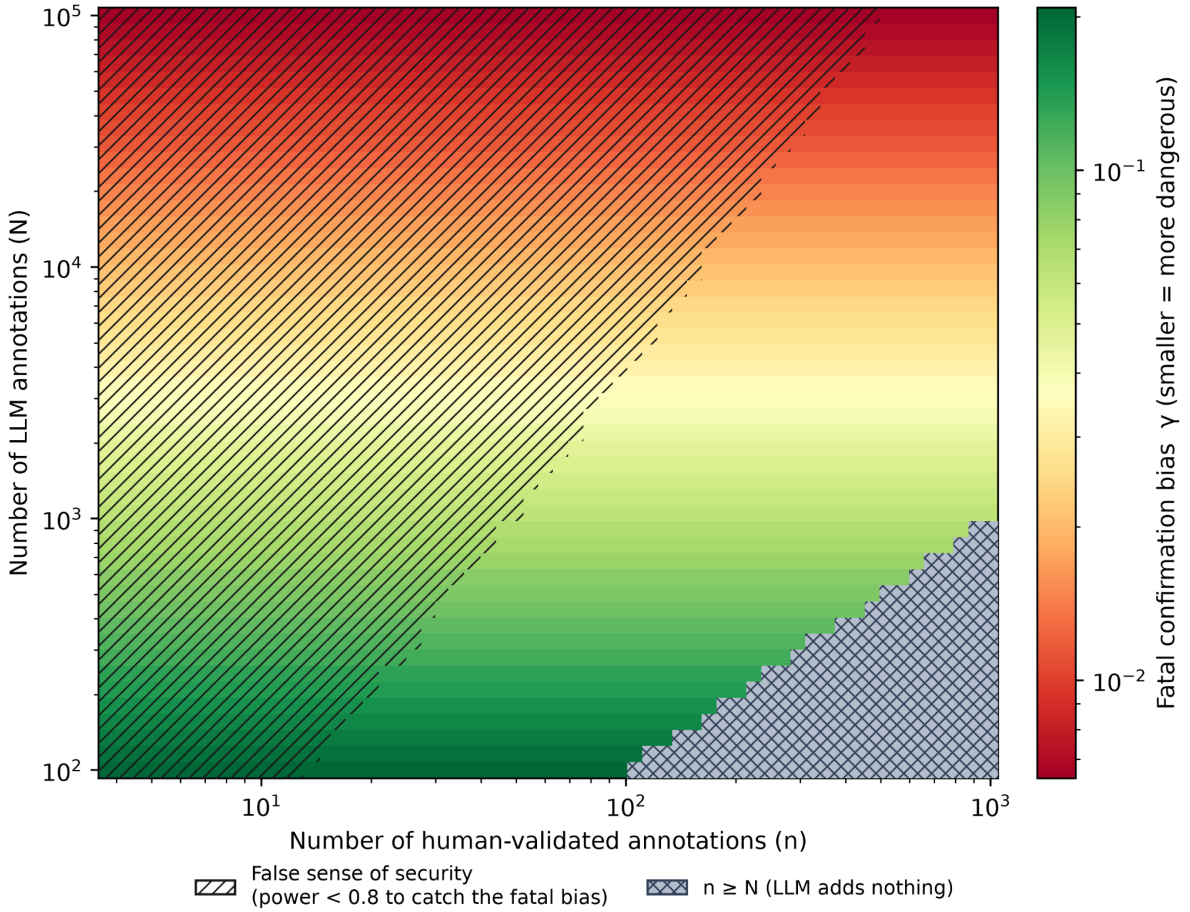

Figure 3. The fatal confirmation bias and the audit needed to catch it. The corpus size N runs up the vertical axis (log) and the audit size n along the horizontal axis (log). Cell colour is γ_{fatal}, the smallest hypothesis-aligned bias that pushes a typical study to p = .05; because it depends on N only, the colour forms horizontal bands, from green (a sizeable bias is needed, N small) to red (a minuscule bias suffices, N large). The hatched region, "False sense of security", marks designs whose audit n is too small to detect γ_{fatal} with 80% power; its frontier rises with N, so the audit must grow with the corpus. Blue cross-hatched cells have n ≥ N.

In the "False sense of security" region (hatched), the audit of n pairs is too small to detect the fatal bias with 80% power: a researcher would run the validation, observe acceptable agreement, feel reassured, and still publish a false positive, because the audit was never large enough to notice the bias. This region covers about 52% of the realistic grid, so the default audit sizes researchers use are, more often than not, too small for the N they work with: a fixed "we hand-checked 50 items" does not become safer as more items are annotated, it becomes more dangerous.

5. Practical recommendations

For false negatives (Study 1), the task is to make sure the design can detect the effect. Estimate the effect size you expect and the sample size you have, locate your study on the map of Figure 2 (or run the protocol for your exact numbers; see below), and read off the required ϱ. Then check, against your audit of n items, that your annotation actually clears it. Two failure modes are worth ruling out first. If your study sits in the "underpowered by design" zone, the fix is more data, not better annotations. If it sits near the n ≥ N diagonal, you are hand-coding so much that

the LLM buys you little. Everywhere else, a surprisingly modest ϱ (often in the .25 to .65 range) is enough.

For false positives (Study 2), conventional validation does not protect you, and there are multiple solutions.

The first is to audit for direction: from the manually annotated pairs, estimate the relationship between the annotation's error (model score minus human score) and the predictor X. Because the fatal bias shrinks as N grows, the audit must be large enough to detect a bias of size $\gamma_{fatal}(N)$ with 80% power, which means n has to rise as N rises. Then, a slope near zero leaves the effect estimate unbiased. A negative slope pushes against the hypothesized direction and can only attenuate the effect. A positive slope pushes in the hypothesized direction and inflates the measured slope with a spurious component (which overall agreement leaves undetected).

If that does not work, because the audit needed to detect $\gamma_{fatal}(N)$ is larger than the number of pairs anyone can hand-code, there is a second solution: reduce N itself. The fatal bias is not a fixed quantity; it grows as the corpus shrinks, because a smaller sample makes the test less sensitive to any spurious slope. So deliberately analyzing fewer annotations, by drawing a random subsample of the corpus, raises the bias the design can tolerate and can pull a study out of the false-positive region without any extra human coding. The two solutions work in opposite directions on the same map: auditing more moves you rightward by raising n, subsampling moves you downward by lowering N, and either one can reach the safe zone.

In our running example, suppose the folklorist has 50,000 tales scored for romance by the model but can only hand-code a few hundred. At that N the fatal bias is tiny, so an audit of a few hundred pairs cannot rule out a slope large enough to manufacture the trend on its own, and the direction test is underpowered. Rather than annotate thousands more tales by hand, she can analyze a random 10,000 instead of all 50,000. The smaller corpus raises the bias she can tolerate, the same audit now suffices to detect it, and the test of whether romance rises over the years becomes trustworthy again, at the cost of some precision she did not need.

Beyond sizing the audit (n) and the corpus (N), we advise preregistering the annotation pipeline (Nosek et al., 2018). The whole Study 2 test rests on the correlation between annotation error and X, and that quantity is only interpretable if the pipeline was fixed before the results were seen. AI annotation involves many choices (the wording of the prompt, the temperature, the model version), and each can change the measured error and its correlation with X. If these choices are made after seeing the outcome, they can be adjusted (even unintentionally) until the error falls below the fatal threshold (Baumann et al., 2025; Alizadeh et al., 2025; Tornberg, 2024). Fixing and recording them in advance is what makes the audit an independent test of the annotator rather than a parameter of the result.

What if the audit is too small to detect the fatal bias and the corpus cannot be shrunk without losing a small but real effect? Here, a preregistered directional hypothesis helps too. We propose that, in this case, when the predicted effect appears, the burden of proof shifts to whoever wants to attribute it to annotation bias rather than to the effect itself: they must supply a hypothesis at least as reasonable and parsimonious as the hypothesized one, explaining why the annotation error would happen to correlate with X in the predicted direction. Suppose we preregister that

romance grows more central in folktales over time, and we find it. To dismiss this as an artifact, a skeptic now has to explain why the model would systematically over-score romance in more recent tales *specifically*. In all, we argue that, in a hypothetico-deductive framework, preregistration also lowers the false positive risk even when the diagnostic cannot be run.

The two thresholds (the annotation quality you need and the audit you must budget) can be computed for your own design. We provide a protocol (the file OSF_SPECIFICATION.md in the project repository: https://osf.io/qdj5s) that helps any researcher through the steps. The file is written so that it can be handed directly to a general-purpose AI assistant, which can then walk the researcher through the procedure and run the simulation on their own design.

For the folktales it would go like this: you state the hypothesis (recent tales are more romantic), the test (a regression of romance on date), the effect size you expect ($\beta = 0.2$), the corpus size ($N = 1{,}000$), the audit you can afford ($n = 100$), and the error you fear. The protocol runs the two simulations described here and returns, for that exact study, the Spearman ϱ your 100-tale audit must show to guard against false negatives (about 0.59 here) and the minimum audit size needed to catch a fatal bias and guard against false positives (here a slope of about 0.064 of the annotation error on X). The entire simulation code is also available on OSF (https://osf.io/qdj5s/).

6. Conclusion

AI annotation is opening the cultural sciences to questions that were unaskable a decade ago, and there are good reasons to welcome it. The field's instinct, to certify an AI annotation once against a universal agreement cutoff and then set the question aside, targets a standard that does not match the inference at stake. There is no annotation quality that is good enough in general; there is only annotation quality that is good enough for a given study, and the study itself determines what that is. Random error hides real effects, and a larger sample compensates for it; systematic error creates false ones, and a larger sample makes it worse (Meng, 2018; Ioannidis, 2005). If researchers hold those two facts apart, audit for the direction of error and not merely its size, and let the human audit grow with the corpus, then the thousand folktales, or the million tweets, will tell them something true.

7. Data Availability

All code and simulation outputs are publicly available on the Open Science Framework at https://osf.io/qdj5s/overview. The study uses no empirical data; all results are produced by deterministic, version-controlled Monte Carlo code that fully reproduces the reported figures and numbers.

8. Acknowledgement

This study was supported by the EUR FrontCog grant ANR-17-EURE-0017 and ANR-10-IDEX-0001-02 to PSL. It has also received support under the Major Research Program of PSL Research University "CultureLab" launched by PSL Research University and implemented by

ANR with the references ANR-10-IDEX-0001. This work was further supported by the ANR MAGIC grant (ANR-25-CE28-5813-01) to E.D.

9. References

Alizadeh, M., et al. (2025). Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. Journal of Computational Social Science, 8, 17.

Bail, C. A. (2024). Can generative AI improve social science? Proceedings of the National Academy of Sciences, 121(21), e2314021121.

Baumann, J., et al. (2025). Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. arXiv:2509.08825.

Baumard, N., Huillery, E., & Zabro, L. (2023). The cultural evolution of love in history. Nature Human Behaviour.

Boelaert, J., Coavoux, S., Ollion, É., Petev, I., & Präg, P. (2025). Machine bias: How do generative language models answer opinion polls? Sociological Methods & Research.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). Measurement Error in Nonlinear Models: A Modern Perspective (2nd ed.). Chapman & Hall/CRC.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Dubourg, E., Thouzeau, V., & Baumard, N. (2024). A step-by-step method for cultural annotation by LLMs. Frontiers in Artificial Intelligence, 7, 1365508.

Dubourg, E., Thouzeau, V., Borredon, Q., & Baumard, N. (2025). Quantifying and explaining the rise of fiction. Evolutionary Human Sciences, 7. https://doi.org/10.1017/ehs.2025.10011

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. Science, 380(6650), 1108–1109.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.

Karjus, A. (2024). Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. arXiv:2309.14379.

Krippendorff, K. (2018). Content Analysis: An Introduction to Its Methodology (4th ed.). SAGE.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics, 12(2), 685–726.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176–182.

Moretti, F. (2000). Conjectures on world literature. New Left Review, 1, 54–68.

Morin, O., & Acerbi, A. (2017). Birth of the cool: A two-centuries decline in emotional expression in Anglophone fiction. Cognition and Emotion, 31(8), 1663–1675.

Nakawake, Y., & Sato, K. (2019). Systematic quantitative analyses reveal the folk-zoological knowledge embedded in folktales. Palgrave Communications, 5, 161.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11), 2600–2606.

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative AI requires validation. arXiv:2306.00176.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. Proceedings of the National Academy of Sciences, 121(34), e2308950121.

Sobchuk, O. (2018). Charting Artistic Evolution: An Essay in Theory (Doctoral dissertation). University of Tartu Press.

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72–101.

Törnberg, P. (2024). Best practices for text annotation with large language models. arXiv:2402.05129.

Underwood, T. (2019). Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? arXiv:2305.03514.