Glossary

The following definitions refer to recurring interpretive patterns within the frequentist approach according to what articulated by Greenland et al. (2016), and are intended to clarify how these terms are used in the present framework.

Subordination - We use this term to denote relegating foundational model assumptions to the cognitive background, treating them as secondary or ignoring them despite their central role in shaping the analysis.

Confidence - We use this term to denote treating interval estimates as if they guaranteed coverage or quantified overall uncertainty, and to misuse them as a plausibility filter separating “plausible” from “implausible” values, while overlooking strong assumptions and ignoring nonstatistical/systematic sources of uncertainty.

Reification - We use this term to denote attributing substantive or evidential status to data or estimated effects, treating them as if they possessed practical meaning or probative force even though they arise from idealized mathematical abstractions.

Simplification - We use this term to denote omitting or distorting interpretively essential details in order to present a simpler statement, thereby making the expression easier to convey but easier to misread.

Significance - We use this term to denote conflating statistical significance with effect importance, and to report results as “significant” or “not significant” without the statistical qualifier, inviting confusion with practical or clinical relevance.

Inversion - We use this term to denote interpreting or describing probabilities defined under the assumption that a model holds as if they instead expressed the probability that the model holds.

Support - We use this term to denote treating compatibility or decision measures as if they provided “support” for hypotheses, even though support is a broader concept and these measures were not designed within a framework that quantifies (evidential) support.

Nullism - We use this term to denote giving the null hypothesis of no effect, no association, or no difference a privileged cognitive or practical status relative to the many alternative hypotheses, at the expense of contextual relevance.

Ritualism - We use this term to denote adherence to institutionalized statistical conventions as automatic rituals, applied without the contextual judgment and problem-specific justification they would require.

old-Dichotomy - We use this term to denote presenting results in a binary way as “significant” versus “not significant,” reducing graded evidence and uncertainty to a two-category label.

neo-Dichotomy - We use this term to denote replacing or mixing the classic significant vs. not-significant split with newer binary labels (e.g., compatible vs. incompatible, convergent vs. divergent), again collapsing a continuum into two categories.

Precision - We use this term to denote misreading inferential labels as statements about precision, instead of assessing precision directly from the spread of interval estimates or the information conveyed by the minimum, statistical uncertainty distribution.

Hybridation - We use this term to denote mixing distinct statistical frameworks, blending compatibility-based uses of P-values with long-run decision procedures, and thereby creating interpretations and practices that are not coherent within either framework.

Comparison - We use this term to denote judging differences between results by comparing the P-values for the tested hypothesis, rather than performing an appropriate direct comparison of the estimates.

Power - We use this term to denote interpreting a non-significant result as favoring the null when power is high, or as disproving the alternative, conflating non-rejection with evidential support.

Confirmation - We use this term to denote selectively focusing on a favorable example or subset of cases to justify a claim that is broadly incorrect, rather than evaluating the totality of relevant evidence.

Operational Coding Rules for Interpretive Patterns

Multiple codes per item are allowed; code all that apply.

Subordination

1. Does the item make an inferential, interpretive, or decision claim about point or interval estimates, P-values, or statistical tests?
2. Does the item **not** state that the evaluation depends on the specific test/model, set of foundational assumptions, or analyst choices?
3. If 1 **and** 2 = yes, code "Subordination".

Confidence

1. Does the item treat an interval estimate (like a "confidence" interval) as guaranteeing coverage or as representing **overall (**random + systematic) uncertainty?
2. Does the item use an interval estimate as a plausibility filter (e.g., to discriminate between "plausible" and "implausible" values)?
3. If **any** of 1 or 2 = yes, code "Confidence".

Reification

1. Does the item talk as if statistical outputs (data, estimates, intervals, P-values) directly reveal or prove a real-world fact (e.g., "the data show/demonstrate that [...]")?
2. Does the item treat an abstract statistical quantity as a property of reality (e.g., "the P-value shows/demonstrates that the effect is significant")?
3. If **any** of 1 or 2 = yes, code "Reification".

Note: common triggers include "show", "demonstrate", "prove", "establish", "reveal" used as if statistical outputs directly certify a real-world fact.

Simplification

1. Does the item replace a technically correct but longer expression with a shorter one that changes the meaning in a misleading way (e.g., "probability of results at least as extreme" -> "probability of the results")?
2. Does the item make the statement easier to grasp but misleading (e.g., "what the test estimates" -> "what the data say")?
3. If **any** of 1 or 2 = yes, code "Simplification".

Significance

1. Does the item use "significant" or "not significant" without explicitly saying "statistically", in a way that can be read as practical/clinical importance?
2. Does the item treat statistical significance as evidence that an effect is important, meaningful, or relevant (or treat non-significance as evidence it is unimportant)?
3. Does the item move from "statistically significant/not significant" to a claim about real-world impact without separating statistical results from practical/clinical relevance?
4. If **any** of 1-2-3 = yes, code "Significance".

Inversion

1. Does the item treat a P-value as "the probability that H_0 is true/false given the data" or that the result is "due to chance"?
2. Does the wording effectively flip "probability of results as or more extreme given the model" into "probability of the model given the data or results"?
3. If **any** of 1 or 2 = yes, code "Inversion".

Support

1. Does the item say that a P-value, interval, or a threshold-based decision "supports" or "provides evidence for" a hypothesis, model, or claim?
2. Does the item use "support" wording without specifying a causal/data-generating framework that defines and justifies "support" (e.g., without clarifying what support means or how it is being quantified)?
3. If **any** of 1 or 2 = yes, code "Support".

Nullism

1. Does the item treat the null (no effect/association/difference) as the default or privileged hypothesis that results must overturn?
2. Does the item focus interpretation mainly on whether the null is rejected/ruled out, rather than on the range of context-relevant alternative hypotheses or effect sizes?
3. Does the item imply that the null deserves special status (e.g., as the main reference point) even when other plausible hypotheses are more relevant?
4. If **any** of 1-2-3 = yes, code "Nullism".

Note: Do **not** code "Nullism" when the prompt/question explicitly asks about the null hypothesis; code it when the item introduces or elevates the null as the focal reference despite a more general or alternative-relevant framing (e.g., the sentence under evaluation talks about a generic “test hypothesis” and the answer shifts the attention or mentions the null).

Ritualism

1. Does the item apply a conventional rule or threshold (e.g., alpha = 0.05, "significant/not significant") as an automatic decision criterion?
2. Does the item treat following the convention as sufficient justification for the conclusion, interpretation, or action?
3. Does the item use the convention without any context-specific rationale (why this rule/threshold fits the problem)?
4. If **any** of 1-2-3 = yes, code "Ritualism".

old-Dichotomy

1. Does the item present results in a binary way as "significant" vs "not significant"?
2. If 1 = yes, code "old-Dichotomy".

neo-Dichotomy

1. Does the item replace or mix the classic "significant/not significant" split with newer binary labels (e.g., "compatible vs. incompatible", "consistent vs. inconsistent" "convergent vs. divergent")?
2. If 1 = yes, code "neo-Dichotomy".

Precision

1. Does the item make a claim about the precision of an estimate/interval?
2. Does it assess precision using a misconception like "includes the null vs excludes the null" (or similar unfounded rules), rather than evaluating interval width/spread?
3. If 1 **and** 2 = yes, code "Precision".

Hybridation

1. Does the item combine neo-Fisherian language (e.g., interpreting P-values and intervals as graded measures of compatibility or statistical evidence) with long-run decision

language (e.g., fixed alpha rules, reject/accept as procedure guarantees) in the same claim?
2. If 1 = yes, code "Hybridation".

Comparison

1. Does the item claim that two results differ because one P-value is smaller/larger than the other?
2. Does it compare P-values from separate tests instead of comparing the estimates directly (e.g., via a contrast/interaction or an appropriate direct test of the difference)?
3. If **any** of 1 or 2 = yes, code "Comparison".

Power

1. Does the item interpret a non-significant result as evidence for the null (or against a meaningful alternative) because the study/test had high power?
2. Does it claim that non-significance "rules out" an effect/alternative, or implies the null is supported, based mainly on power considerations?
3. If **any** of 1 or 2 = yes, code "Power".

Confirmation

1. Does the item justify a broad claim by highlighting a favorable example or subset while downplaying or ignoring the rest of the relevant evidence?
2. Does the comment hinge on selective emphasis rather than an overall assessment?
3. If **any** of 1 or 2 = yes, code "Confirmation".

Inter-rater Agreement Results (Pre-comparison Coding)

Inter-rater agreement between two raters was evaluated on 18 interpretive statements using pre-comparison multi-label coding. Statements were randomly sampled from real outputs generated by the large language models under study. Given the multi-label nature of the coding scheme, agreement was quantified using set-based similarity indices rather than single-label kappa.

Overall agreement

Micro-averaged Jaccard index = 0.816, 95% CI: 0.709 - 0.912

Micro-averaged Dice coefficient = 0.899, 95% CI: 0.830 - 0.954

Item-level Agreement (Macro Averages)

Macro-averaged Jaccard index = 0.838, 95% CI: 0.742 - 0.921

Macro-averaged Dice coefficient = 0.898, 95% CI: 0.830 - 0.954

Notes

Agreement was assessed before any discussion or reconciliation, reflecting genuine independent judgment. Confidence intervals were obtained via bootstrap resampling over items. Traditional Cohen's kappa was not used because the coding scheme is multi-label, and kappa assumes a single categorical label per item. Set-based indices (Jaccard, Dice) are more appropriate in this context.

Concrete examples of fallacies

Subordination

The CI describes the uncertainty of the mean - Gemini 3

Statistical significance is not a property of the phenomenon; it's a property of the data relative to a chosen threshold - Copilot

Notes: the backgrounding of key assumptions recurs across most statements; the examples below therefore suffice to illustrate it (e.g., see the "Confidence" fallacy section).

Confidence

A 95% confidence interval does not give a 95% probability for the true effect; it means if you repeated the study many times, 95% of such intervals would contain the true effect - ChatGPT 5 mini

A specific 95 % confidence interval does not have a 95 % probability of containing the true effect; the method has 95 % long-run coverage - ChatGPT 5.2

Once the CI is calculated from data, it either contains the true parameter or it doesn't—it's fixed, so the probability is 0% or 100%. The 95% refers to the long-run coverage probability of the method: 95% of CIs from repeated samples will contain the true value - Grok 4

The 95% refers to the process: if we did the study 100 times, 95 of the resulting intervals would contain the truth - Gemini 3

In repeated samples, 95% of such intervals would contain the true effect - Copilot

Reification

Significance is a function of data, sample size, and variability - Perplexity

A small p-value increases confidence in a real effect - Grok 4

Statistical significance is a property of the data - ChatGPT 5 mini

Confidence interval = range of plausible values given the data - Copilot

By definition, a 95% confidence interval excludes effect sizes outside it with 95% confidence, meaning the data refute (or rule out) those values as plausible for the true effect at the 5% significance level - Grok 4

Simplification

A P-value quantifies the probability of obtaining data at least as extreme as observed under the null. Saying "chance alone produced the observed association" is roughly correct - ChatGPT 5 mini

"P-values are properly reported as inequalities (e.g., report "P < 0.02" when P = 0.015 or report "P > 0.05" when P = 0.06 or P = 0.70)." This is acceptable and common practice in some fields (e.g., psychology, medicine) to avoid overprecision or when exact values aren't computed [...] - Grok 4

The P-value is the probability of observing data as extreme or more extreme than those observed, assuming the test hypothesis is true. The phrasing "only 5 % of the time" is informal but essentially correct. - ChatGPT 5.2

"The P-value is the chance of our data occurring if the test hypothesis is true." Correct. This matches the definition: probability of data (or more extreme) if H0 true - Perplexity

P = 0.05 means exactly 0.05; P ≤ 0.05 includes values up to 0.05. Inequalities like P < 0.02 for P=0.015 are standard and precise - Copilot

Reporting as inequalities (e.g., P < 0.02 for 0.015) is a recommended practice for precision. - Perplexity

Significance

A significant result (P ≤ 0.05) means there is evidence against the null hypothesis - Perplexity

Two estimates can have overlapping 95% CIs and still be significantly different from each other - Gemini 3

Overlapping confidence intervals do not guarantee non-significance - ChatGPT 5 mini

Inversion

Statistical significance indicates that an observed relation is unlikely to be due to chance alone - Grok 4

If the P-value is small (e.g., P ≤ 0.05), we conclude that the observed data is very "unlikely" to have happened by coincidence alone, which gives us a reason to reject the null hypothesis - Gemini 3

“The P-value for the null hypothesis is the probability that chance alone produced the observed association; [...]” Correct. The P-value represents the probability that the observed association (or more extreme) occurred by chance alone if the null hypothesis is true, so P=0.08 indicates an 8% chance of that happening due to chance. - Perplexity [Contradictory]

Support

Evidence for H_0 requires methods like equivalence testing or high power with nonsignificance - Grok 4

P > 0.05 with high power suggests evidence favoring the null - Perplexity

Nullism

Q9 is correct (with minor phrasing note: "test hypothesis" typically means null). - Perplexity

The P-value is the probability of observing the data if the null hypothesis is true (note: “test hypothesis” usually means null here) - ChatGPT 5 mini

Ritualism

Correct: Reporting P-values as inequalities is standard practice - ChatGPT 5 mini

Inequalities like "p > 0.05" are fine for nonsignificant results - Grok 4

P = 0.05 and P ≤ 0.05 are treated the same in hypothesis testing - Copilot

old-Dichotomy

P > 0.05 means no evidence against the null - Copilot

A nonsignificant result (P > 0.05) means insufficient evidence to reject the null - Perplexity

If the P-value exceeds 0.05, the data do not provide strong evidence against the null hypothesis - ChatGPT 5.2

Notes: see also the examples above and below (e.g., see the “Ritualism” fallacy section or the “Comparison” fallacy section).

neo-Dichotomy & Hybridation

An effect outside a confidence interval is not “refuted” with absolute certainty; it is simply incompatible with the data at the given confidence level - ChatGPT 5 mini

Never use strong words like "refuted," "disproven," or "excluded" based on a confidence interval. Instead, say that those values are "inconsistent with the data at the 0.05 level" - Gemini 3

Exclusion at 95% level means not compatible, but not fully refuted without power consideration - Perplexity

A P-value > 0.05 means the data are compatible with no effect - Perplexity

Statistical significance is about compatibility with a model. Lack of statistical significance does not imply a small effect - ChatGPT 5.2

Precision

Notes: no examples.

Comparison

Opposite sides of 0.05 suggest conflicting results regarding statistical significance - ChatGPT 5 mini

Opposite sides of 0.05 suggest conflicting evidence - Copilot

Same P-values in different populations indicate agreement in terms of significance - ChatGPT 5 mini

If none or a minority of studies are significant (all P > 0.05), the overall evidence (e.g., via meta-analysis) typically does not support the research hypothesis (alternative) - Grok 4

Power

If the test has 90% power at a meaningful alternative, and the null P-value is >0.05, the data provide more support for the null than for that specific alternative - ChatGPT 5 mini

Nonsignificance with high power suggests evidence for null - Perplexity

A high-power non-significant result suggests the effect is not as large as the one you powered for - Gemini 3

P > 0.05 with high power suggests evidence for the null over the alternative - Copilot

Confirmation

Reporting as inequalities is standard (e.g., "P < 0.001") - Copilot

Correct. Exact P-values are preferred, but inequalities are common and acceptable for small or large values - Perplexity

"P = 0.05 is equivalent to P ≤ 0.05." ✅ Correct – By definition, if P = 0.05, it also satisfies P ≤ 0.05. - ChatGPT 5 mini

Threshold-Based Coding of Expanded Outcomes

The expanded outcomes rely on interpretive coding of natural-language explanations. Unlike the primary endorsement/rejection outcome, which is relatively straightforward to operationalize, the textual fallacy coding involves judgment calls about whether a given response instance matches one of several predefined interpretive patterns. Accordingly, we summarize expanded outcomes using coarse, threshold-based categories rather than fine-grained percentages.

Meaningful category-level prevalence estimates would require conditioning on item-level "opportunities" for each fallacy category (i.e., specifying which benchmark statements plausibly afford which fallacies). Because benchmark statements differ substantially in content (e.g., only a subset concerns confidence intervals or power), unconditioned percentages can conflate model behavior with exposure to the relevant context. Defining eligibility sets and estimating P(fallacy | eligible) would add further analytic layers and assumptions, and would shift emphasis toward modeling choices that are themselves partly interpretive.

In addition, interpretive coding of free-form explanations is inherently probabilistic: even with satisfactory inter-rater agreement, coders share background assumptions about statistical reasoning, and linguistic realizations of the same fallacy can vary widely across responses. Finally, because LLM behavior can vary across runs and change over time due to updates and interface-level factors, the study is framed as characterizing educational risk under realistic usage conditions rather than estimating stable intrinsic properties of a model.

For these reasons, we adopt a risk-oriented, semi-quantitative representation: for each glossary category we report whether it was never observed ("No"), observed sporadically ("Occasional"), or observed repeatedly ("Yes") across the 26 primary runs. Specifically, we define "No" as 0/26, "Occasional" as 1-3/26, and "Yes" as greater than or equal to 4/26. These thresholds are intended to support transparent risk detection and communication (presence and repeatability of misleading language) rather than fine-grained prevalence estimation.

References

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology, 31(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. The American Statistician, 73(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Rovetta, A., & Mansournia, M. A. (2025). Common wrong beliefs about statistical testing: Recent trends in biomedical sciences. *Current Opinion in Epidemiology and Public Health, 4(3), 23-29. https://doi.org/10.1097/PXH.0000000000000054