From Detection to Correction: A Hybrid NLP Approach to Misinterpretations of Nonsignificant *p* Values

Raphael Merz

Department of Psychology, Ruhr University Bochum

Author Note

Raphael Merz https://orcid.org/0000-0002-9474-3379

Correspondence concerning this article should be addressed to Raphael Merz, Email: raphael.merz@rub.de

Abstract

Misinterpretations of p values remain widespread in scientific reporting, despite decades of educational efforts and reform initiatives. One of the most common and consequential errors is interpreting a statistically nonsignificant result (e.g., p > .05) as evidence for the absence of an effect—a conclusion not supported by null hypothesis significance testing (NHST). This thesis adopts a human factors perspective, arguing that automation can help mitigate such persistent errors, much like word processors assist with grammar and spelling. I propose an automated, three-step pipeline that detects, classifies, and optionally corrects misinterpretations of nonsignificant results. Evaluation of each step highlights the promise of such automated approaches: In a validation set of 25 articles, the automatic detection identified 73% of human-extracted statements. Two easily resolvable issues in the search pattern were found which, once addressed, would increase this reliability to 93%. For classification, three BERT-based models were trained on 930 hand-labeled statements. All performed well, with SciBERT achieving the highest macro F1 score of .91. Finally, the optional correction step proved effective in a validation set of 80 incorrect and 20 correcct statements: 85 statements were correctly phrased after LLM-based revision. These results demonstrate that automation can effectively address this specific misinterpretation and offer a flexible foundation for tackling similar issues in scientific writing and meta-research.

Keywords: p value, misinterpretation, automation, automated checks, RegEx, LLMs, BERT

From Detection to Correction: A Hybrid NLP Approach to Misinterpretations of Nonsignificant *p* Values

1 Introduction

Over the past decades, numerous articles have addressed common misinterpretations of *p* values in the context of standard null hypothesis significance testing (NHST) (Goodman, 2008; Greenland et al., 2016; Schervish, 1996). Some go further, questioning the use of frequentist methods altogether (Edwards et al., 1963; Wagenmakers, 2007), while others propose refinements within the frequentist framework that aim to improve the informativeness of statistical inference (Isager & Fitzgerald, 2025; Lakens et al., 2018). If you are a researcher writing a paper and want to interpret your results correctly, the solution seems simple: read these educational resources and revise your manuscript accordingly. Easy, right? Still, empirical studies consistently show that these misinterpretations remain widespread (Hoekstra et al., 2006; Murphy et al., 2025). Why is that? What makes interpreting *p* values so persistently difficult? Which practical solutions or promising approaches might help? And are some of the proposed 'misinterpretation checklists' perhaps less informative than their authors would hope?

In this article, I show how rule-based approaches, combined with natural language processing (NLP) can be used to automatically detect, classify, and correct these misinterpretation. I show this for the misinterpretation of statistically nonsignificant results as the absence of an effect because it is the most extensively researched misinterpretation of *p* values (Lakens, 2021), and I have experience in classifying them from a previous project (Murphy et al., 2025). That said, the general framework I propose can be easily adapted to address other misinterpretations beyond *p* values.

1.1 Misinterpretations and Criticism of *P* Values

The criticism of *p* values has become a prominent and recurring theme in discussions around scientific reform. From claims that they encourage dichotomous thinking (Amrhein et al., 2019; Hoekstra et al., 2006) to arguments that they offer little informational value (Wagenmakers, 2007), *p* values – and the broader framework of NHST – have been blamed for many of science's replication problems (McShane et al., 2019). On the other hand, many have also argued that NHST per se is not to blame for these problems, but rather how researchers (mis)use and (mis)interpret this tool (Lakens, 2021). As a result, many researchers present whole collections of, in their view, common *p* value misinterpretations (see, e.g., Goodman, 2008; Greenland et al., 2016). Reviewing these, I see four key misconceptions about *p* values that seem to be at play:

- *p* values as hypothesis probabilities
- *p* values as measures of practical significance
- *p* values as measures of replicability or error rates
- Technical misunderstandings about *p* values

[NOTE FROM RAPHAEL: I left this part in for now, because it was part of the original proposal and I didn't fully rewrite the Intro yet: However, I think that this doesn't really work for the thesis/this article anymore. I feel like this would be a Review paper on its own and then automatically fixing some of these misconceptions would be, again, a new paper. Given the tight word limit (5000 words), I think I will just briefly put the "p > .05 = no effect" misinterpretation in the context of p value criticism in general, but then not mention this "I came up with these four categories" bit. But let me know what you think!]

In my master thesis, I will zoom in on one specific misinterpretation: concluding *no effect* based on a statistically nonsignificant finding. Many studies have previously shown that this misinterpretation is and remains highly prevalent across time and sub-domains of psychology (e.g., Aczel et al., 2018; Hoekstra et al., 2006; Murphy et al., 2025). In fact, in a recently published article investigating articles published in 2009, 2015, and 2021 across ten different psychology journals, we estimated the prevalence of this misinterpretation in articles' discussion sections to lie between 76.17% and 84.90% (Murphy et al., 2025). This study highlights that the situation seems not to have greatly improved despite many researchers exploring new analysis techniques (e.g., Lakens et al., 2018) and continuous calls to reflect on interpretations of nonsignificant results (e.g., McShane et al., 2019).

1.2 Possible Solutions

One frequently suggested solution is to improve researchers' statistical literacy through enhanced education, such as better statistics teaching at the undergraduate and graduate levels (e.g., Lakens, 2021). However, as noted earlier, the prevalence of the misinterpretation I focus on does not seem to have substantially decreased, suggesting that calls for better education alone have not resolved the problem (Murphy et al., 2025). Relatedly, researchers have also advocated for the use of interval hypotheses tests, like equivalence testing or minimum-effect tests (or the combination: three-sided testing; Isager & Fitzgerald, 2025). These methods allow researchers to test whether an effect is practically relevant and larger than a predefined smallest effect size of interest (SESOI; Lakens et al., 2018). In many contexts, such approaches might be more closely aligned with the substantive questions researchers aim to answer, namely whether an effect is meaningful in practice.

These strategies also align with the argument made by Lakens (2021) that *p* value misinterpretations represent a human factors problem, requiring practical and easy-to-implement solutions. In other context, we encounter systems like this frequently, be it cars' automatic braking systems, word processors that flag spelling and grammar mistakes, or email clients that filter out malware and phishing attempts. Analogously, and recognizing that new analytic approaches may not be adopted overnight, automated checks for statistical misinterpretations offer a highly promising route. This perspective emphasizes that many statistical errors arise not from bad intentions or ignorance, but from cognitive limitations and suboptimal workflows.

In the context of research, similar automated solutions are already gaining traction. For instance, the reference manager Zotero flags references to retracted papers (Stillman, 2019). Statcheck (Nuijten & Epskamp, 2024) automatically detects inconsistencies between reported test statistics and *p* values. Other tools, like GRIM, GRIMMER, and SPRITE, identify impossible values in reported summary statistics (Heathers et al., 2018). And lastly, Regcheck (Cummin & Hussey, 2025) verifies the consistency between manuscripts and their preregistration documents.

To make the process of checking manuscripts more systematic, DeBruine and Lakens (2025) developed papercheck, an R package, which allows users to run a battery of checks on scientific papers. These include statistical checks (e.g., identifying imprecisely reported *p* values) as well as general manuscript quality checks (e.g., verifying links to online repositories or consistency between in-text citations and reference lists). Papercheck can be used both for single articles (e.g., as writing assistance) and for batches of articles (e.g., for meta-scientific studies). Because this framework is actively maintained and continues to evolve, the approach presented in this study was designed to fit within the papercheck infrastructure.

In summary, there are many reasons why *p* values remain difficult to interpret correctly. Empirical evidence suggests that misinterpretations of nonsignificant results remain highly prevalent (Murphy et al., 2025). This persistence highlights that improved education alone may not be sufficient. Drawing on a human factors perspective (Lakens, 2021), practical solutions such as automated error-checking tools offer a promising avenue for addressing these challenges. In this project, I develop a pipeline to automatically detect, classify and optionally correct one such error: interpreting nonsignificant results as the absence of an effect. The pipeline builds on and is integrated into the existing papercheck framework (DeBruine & Lakens, 2025), a modular R package designed to support statistical and quality checks across scientific manuscripts. The next section outlines the methods I use to implement and evaluate this approach.

2 Methods

2.1 Statement Detection, Classification and Correction

Before describing the data used in this study, it is important to understand the three steps of the proposed framework. Statements from scientific articles needed to be reliably detected, classified, and finally corrected. For each step, I applied specific methods that were best suited to achieve the respective goal.

To detect statements I searched used rule-based regular expressions (RegEx) and searched articles's results sections for them. Effectively, RegEx searchers are more complex Ctrl+F searches, where a user can also include rules like optional characters (e.g., ‘significant(ly)’ would catch both *significant* and *significantly*) and more complex rules (e.g., ‘not.{0,20}significant’ allows up to 20 characters between *not* and *significant*). Papercheck (DeBruine & Lakens, 2025) has a module that detects almost all *p* values (see Section 3.1 for examples of what it does not currently detect) based on RegEx searches and I filtered these to just the ones equal to or above .05. I then expanded the extracted nonsignificant *p* values to the full sentence with papercheck and added +/- one sentence as context in case of extraction errors (incomplete statements).[1]

In the next step, these statements (labeled as correct or incorrect by me; see Section 2.2) were used to train three BERT-based models. BERT (Bidirectional Encoder Representations from Transformers) is a general-purpose language model pre-trained on the BookCorpus and English Wikipedia, making it suitable for a wide range of tasks – but not specifically optimized for scientific or technical language (Devlin et al., 2019). Since its introduction, many researchers have developed domain-specific variants of BERT to enhance its performance on specialized tasks. To test whether such domain adaptation improves performance in this study's classification task, I trained two models in addition to standard BERT: SciBERT was trained on a large corpus of scientific articles from Semantic Scholar, particularly in the biomedical and computer science domains (Beltagy et al., 2019). PubMedBERT goes even further, having been trained exclusively on biomedical abstracts and full-text articles from the PubMed database (Gu et al., 2022). These models were trained and evaluated on their ability to distinguish between correct and incorrect (as the absence of an effect) interpretations of nonsignificant results in scientific writing.

Lastly, statements that were classified by the best performing BERT model, would, in the application of this framework, be sent to an LLM to be corrected. However, for the purposes of this validation study, I sent statements which I coded as correct on incorrect to the LLM (more on this in Section 2.2). The full prompt is available in X was had X word. In short, the model was instructed to only change any misinterpretations of nonsignificant *p* values as the absence of an effect and keep the rest of the statement unchanged. To communicate with the LLM, I used papercheck (DeBruine & Lakens, 2025), which, in turn, uses the Groq API (available at https://groq.com/) to communicate with different LLMs. I used papercheck's standard LLM, 'llama-3.3-70b-versatile' (as of 07/24/2025).

[1] In a final tool, users will be able to set the alpha level they used themselves, thus allowing other levels than the conventional 5%.

2.2 Validation Process and Performance Metrics

To assess how well each of these three automated approaches worked, I compared each one to human ground truth and calculated appropriate measures of reliability between automated and human results.

Firstly, to ensure that the statement detection process actually caught all statements with nonsignificant *p* values in articles' results sections, I manually extracted these from 25 (10%) of the papercheck sample library's 250 open access article from the journal Psychological Science. These articles were published between 2013 and 2024 (Median = 2021; IQA = [2018; 2022]). I then coded whether a statements I found were also extracted with the automated RegEx search.

For the training of the BERT models and to assess their final performance, I labeled all automatically extracted statements that were detected to be from an article's results section from papercheck's sample library. This resulted in 960 statements in total. Of these, 419 were classified as correct and 353 were classified as incorrect. The remaining 188 statements were classified as neither correct nor incorrect because they interpreted the nonsignificant effect as (marginally) significant (83), because the statements were not complete enough to check their correctness (20), because they interpreted model fit indices and not the *p* value (20), because they were 'false flags' of nonsignificant *p* values (19), nor really containing nonsignificant *p* values, or due to a combination of these or other reasons (46).

Lastly, I also also went through the 100 statements that were sent to an LLM to be corrected, to see if the 'corrected' statements were actually correct. Of these 100 statements, 80 had previously been labeled incorrect and 20 correct by me, to check how the LLM deals with false positives from the automated classification.

In addition to these validity checks, there are also performance metrics specific to the trained classifiers. For training purposes, the labeled data was split into three parts: a test set (20%) used for the final evaluation of the best model, a training set (72%, or 90% of the remaining 80%) that the model uses to learn underlying patterns and adjust its parameters, and a validation set (8%, or 10% of the 80%) used to calculate evaluation metrics after each epoch (i.e., one full cycle of the model processing the training data). Before the data was split into these three parts, I balanced the two labels (originally there were more correct than incorrect interpretations) to ensure that the model would not overfit to this class-imbalance.

During BERT training, I computed the training loss (sum of errors between model predictions and actual labels in the training set) and the validation loss (same for validation set). The best-performing model was selected based on the lowest validation loss. The model would have been trained on a maximum of 16 epochs, but training ended early if the model did not improve, as measured by the validation loss, for two epochs. In fact, the longest training was 12. For final evaluation, I computed the fraction of correctly predicted classes among all predicted cases of a class (precision), the fraction of correctly predicted classes among all actual cases of a class (recall), and their harmonic mean (F1 score), separately for each class. To summarize overall performance across the two classes, I calculated the unweighted average of the two F1 scores (macro-F1 score).

2.3 Software

All scripts for this thesis project were written in R [Version 4.5.0; Team (2025)] or Python [Version 3.12.10; Python Software Foundation (2025)].

In R, I used *papercheck* [Version 0.0.0.9049; DeBruine and Lakens (2025)] for accessing the 250 open access articles, preprocess them and for communication with the LLM, *readxl* [Version 1.4.5; Wickham and Bryan (2025)] to access Excel files in R, *psych* [Version 2.5.6; William Revelle (2025)] for calculating descriptive statistics, *tidyverse* [Version 2.0.0; Wickham et al. (2019)] for data preprocessing and visualization, and *flextable* [Version 0.9.9; Gohel and Skintzos (2025)], *magick* [Version 2.8.7; Jeroen (2025)], *papaja* [Version 0.1.3; Aust and Barth (2024)] and *showtext* [Version 0.9-7; Qiu and details. (2024)] to create APA-formatted tables.

The Python libraries used to train the BERT models can be found in this requirements file: LINK. [NOTE: They were just to many and apparently it's less common to cite these compared to R packages, which is why its sometimes very tricky to actually get info on how to cite them. I'll look into it later.]

All scripts and data to reproduce and use the trained BERT models (Python), analyse the results and validity checks (mostly R) and recreate this manuscript (Quarto Markdown in

R Studio; I did change few formatting, not content, related things manually in the exported Word document) are available in this GitHub repository, together with instructions on how to set it up: LINK.

This thesis was not preregistered as no inferential statistical tests were performed.

3 Results

3.1 Detection Accuracy

Manually going through 25 articles from papercheck's sample library I detected 179 statements with a nonsignificant *p* value in total. The automated RegEx search caught 130 (73 %) of these completely, and 6 partially (incompletely) due to extraction errors (e.g., because of pdf formatting like page breaks, figures or footnotes). It also 'found' 3 false positives in the sense that it extracted 'statements' from tables or figure notes or ones that were misclassified as coming from a results section. Note, however, that the large majority of the total 49 missed statements were due to specific ways of writing (or not writing) the *p* value: 8 were missed because the authors wrote 'n.s.' instead of the nonsignificant *p* value, and 31 were missed because the *p* value was written as 'p_s'. Without these two mistakes the overall agreement of automated and manual approach would have been 93 %.

Most of the other misses were due to pdf formatting issues like figures, tables, footnotes and page breaks or unusual characters inside the statement that interfered with the statement extraction (11 in total).[2]

3.2 Classification Performance

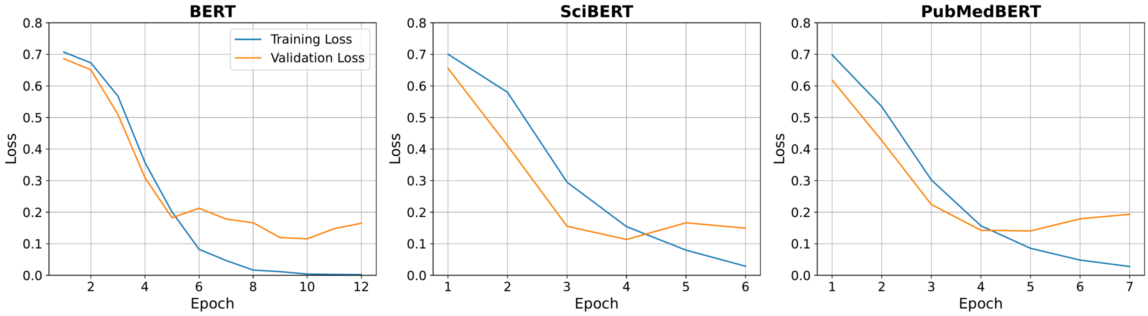
Figure 1 shows the training and validation loss curves for the three BERT models across their training. Standard BERT was trained for a total of 12 epochs before early stopping was triggered due to a lack of improvement in validation loss for two consecutive epochs. The model from epoch 10 was therefore selected as the best-performing version. Similarly, SciBERT and PubMedBERT achieved their best validation performance after epoch 4 and 5, respectively.

[2] I could not find one statement that was extracted automatically in the artilce's pdf. My current theory is that this was an artifact from when the pdf was compiled and might be from a different article even, once again highlighting how impractical the pdf format is in times of increasing automation.

As shown in Figure 1, the training loss consistently decreased over time for all three models, as expected given that they were optimized to fit the training data. In contrast, the validation loss plateaued in all models at a certain point, indicating that further improvements in fitting the training data no longer translated into better performance on unseen data and may even signal the onset of overfitting. Notably, this plateau occurred later in the training of the standard BERT model, which may reflect its different pretraining on general English text compared to the domain-specific pretraining of SciBERT and PubMedBERT.

Figure 1

Training and Validation Loss Curve



Note. Curves of the training and validation loss of the three trained BERT models. The best models for regular BERT, SciBERT and PubMedBERT were chosen after epoch 10, 4, and 5, respectively, based on the minumum validation loss.

The different performance metrics are summarized in Table 1. As reflected in the macro F1 score, both SciBERT achieved the best overall performance in classifying correct and incorrect statements with a macro F1 score of .91. Standard BERT and PubMedBERT lagged slightly behind, with macro F1 scores of .89 and .87, respectively. All three models appear to be better at predicting correct statements then incorrect ones, reflected by the F1 score of the ‘correct’ class, with SciBERT scoring best (‘correct’ F1 score of .91). Similarly, in all models, recall was higher than precision in the ‘correct’ class, whereas the opposite pattern was visible in the ‘incorrect’ class, suggesting that the models tend to err on the side of overidentifying statements as correct rather than incorrect. In fact, the standard BERT model was just slightly better at reducing false negatives (at the cost of more false positives) in this test set (Precision in the ‘incorrect’ class: .91 vs. SciBERT’s .97). Figure 2 reveals more of these sometimes subtle differences between the three models.

Table 1

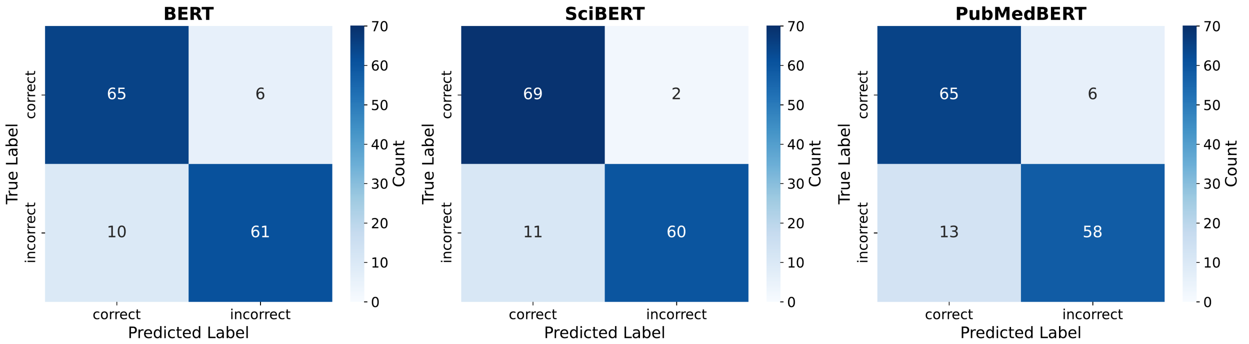
Model Performance

	BERT			SciBERT			PubMedBERT		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
Correct Class	.87	.92	.89	.86	.97	.91	.83	.92	.87
Incorrect Class	.91	.86	.88	.97	.85	.90	.91	.82	.86
Macro F1 score			.89			.91			.87

Note. Table of precision, recall and F1 score per model and class.

Figure 2

Confusion Matrix



Note. Confusion matrices of the three trained BERT models.

Since SciBERT was the best, overall model, Table 2 shows some of its incorrect predictions.

3.3 Correction Evaluation

Of the 100 statements that the LLM was instructed to correct 85 were correct. Interestingly, 2 of the 20 already correct statements got turned incorrect by the LLM. 18, on the other hand, remained correct. Similarly, the LLM actually corrected 67 of the 80 incorrect statements, whereas 13 remained incorrect. Note, however, that the LLM was instructed to change (as much as necessary, but) as little as possible about the original statement. For some statements, this meant that they could not be corrected without major rephrasing. Examples for some bad/good corrections can be found in Table 3 and Table 4, respectively.

4 Discussion

4.1 Summary of Key Results

In this study, I developed and evaluated a three-step pipeline for automatically correcting misinterpretations of nonsignificant results as evidence for the absence of an effect. The approach combines rule-based RegEx searches for detecting candidate statements, fine-tuned BERT models for classification, and LLMs to generate corrected phrasings. While each step leaves room for improvement, the pipeline performed well and shows promise for broader applications beyond *p* value misinterpretations.

Crucially, the framework works effectively because each step is tailored to a specific subtask in the correction process. The RegEx-based detection offers a fast, systematic, and transparent way of reducing the volume of text needing NLP-based analysis. The BERT models provide a lightweight yet powerful solution for learning subtle language patterns. Finally, the LLM correction, while optional, enhances the user experience by offering useful rewording suggestions tailored to a specific correction task. This layered, hybrid structure makes the approach both flexible and scalable.

The RegEx-based statement detection phase demonstrated that simple, rule-based searches can effectively flag a large proportion of candidate interpretations. However, as the study revealed, several limitations remain. For example, statements that reported nonsignificance in a slightly different way (e.g., as "p = n.s." or with subscripted 'p_s') were missed. In some cases, odd formatting issues in pdfs made correct extraction impossible. These limitations highlight the need to continue refining the RegEx patterns and to explore better text extraction techniques.

Still, the classification results are particularly promising given the relatively small size of the manually labeled dataset (< 1,000 examples, split into training, validation, and test sets). The strong performance likely reflects a certain regularity in how nonsignificance is (mis)interpreted in academic writing - commonly through the use of either 'significant' or 'no effect' (e.g., 'there was no effect', but also 'groups did not differ') terminology. While the training dataset was limited to statements extracted from Psychological Science articles (using the existing papercheck sample library), the results provide a solid baseline for expansion using more diverse sources and research domains.

The final correction step, generating corrections of the original statements, was the least developed but offers clear potential. The correction mechanism is optional, and the system performs well even without it. However, when prompted with a clear and narrow task, the LLaMA model used here provided helpful and context-sensitive rewriting suggestions. These suggestions can help users reframe statements in more statistically appropriate ways and reflect on their misinterpretations.

4.2 Limitations and Challenges

Despite the encouraging results, several limitations must be acknowledged. First, the pipeline components were evaluated independently rather than as a fully integrated system. While each step (detection, classification, correction) showed strong performance on its own, cascading errors in a full pipeline will likely reduce the end accuracy. This limitation must be addressed before the framework can be shared.

Second, the manual annotation of training data introduces inevitable subjectivity. While I made efforts to standardize labels - often consulting a statistics expert (my supervisor) on difficult or borderline cases - the classifications reflect my interpretation of what constitutes a misinterpretation. A reader might disagree with how some example statements in this article were labeled, and that is valid. Ideally, multiple annotators and inter-rater agreement metrics would improve the reliability and generalizability of the training data. However, the fact that the fine-tuned BERT models generalized well to unseen data suggests that the labeling was systematic and allowed the models to learn.

A more practical challenge involves managing the tradeoff between false positives and false negatives. The current models aim to balance both for optimal macro performance. However, in practice, different use cases may prioritize one over the other. For example, an individual researcher using the system to improve their writing may prefer fewer false negatives (i.e., catching as many problematic statements as possible), even at the cost of some false positives. Conversely, a meta-scientist analyzing prevalence trends of this misinterpretation may prioritize precision to avoid overestimating misinterpretations. This issue can be mitigated by allowing users to adjust the model's decision threshold to fit their specific goals, something that will also be included in the framework's rollout as part of a papercheck module.

Another limitation involves the narrow context in which statements are classified (a single sentence containing a nonsignificant *p* value). This limited scope means the model cannot account for broader contextual factors, such as whether authors conducted equivalence testing, reported Bayesian results, or provided qualifying language elsewhere. As noted earlier, this study provides a baseline approach intended for future expansion.

4.3 Practical Use and Future Directions

The pipeline described in this study will be integrated into the papercheck infrastructure as a new module for identifying potential misinterpretations of nonsignificant results. While not perfect, the current system is already useful and planned improvements like the ones discussed above will continue to increase its reliability.

Importantly, the step-wise structure of the approach makes it easy to adapt for other classification or correction tasks. For example, users could build their own custom classifiers to detect different types of reporting issues (see van Abkoude, 2025, for an example of using BERT classifiers to classify different problematic use of causal language). In the context of meta science, these classifiers could also be trained with already collected, hand-labeled data (e.g., Aczel et al., 2018).

Going forward, a key step in improving the pipeline's general usefulness will be expanding both the context considered and the scope of what is being searched. In its current version, only single sentences from results sections that contain nonsignificant *p* values are analyzed. The long-term goal, however, is to evaluate entire passages or paragraphs related to a nonsignificant finding in the entire article, in order to capture the full context and provide more meaningful corrections. Achieving this will be essential for addressing the misinterpretation problem more comprehensively.

5 Conclusion

This study demonstrates that a hybrid rule-based and NLP-driven pipeline can effectively detect, classify, and correct a common statistical misinterpretation in scientific writing: interpreting nonsignificant results as evidence for the absence of an effect. Each step - statement detection, classification, and correction - performed well independently. The next step is to evaluate the pipeline as a fully automated system in real-world use cases. With further refinement, this framework has the potential to enhance both automated manuscript checks and large-scale meta-scientific analyses at scale.

Achknowledgement

I want to thank Dr. Daniël Lakens for his constant support throughout this thesis and for a wonderful research stay that allowed me to work on it in person. I thank my partner for her unwavering support, for encouraging me when I felt discouraged, and for our insightful discussions on whether and how we (should and) should not use AI. I also thank Prof. Dr. Maike Luhmann for allowing me to pursue a meta-scientific project that is so close to my heart, even though it falls somewhat outside her area of expertise. Finally, I thank Christian Sodano for helpful discussions on machine learning and BERT models, which helped ensure that my model training did not turn into 'algorithmic *p*-hacking'.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., Van Den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366. https://doi.org/10.1177/2515245918773742

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown* [Manual]. https://doi.org/10.32614/CRAN.package.papaja

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. https://doi.org/10.48550/ARXIV.1903.10676

Cummin, J., & Hussey, I. (2025). *RegCheck. Compare preregistrations with papers. Instantly.* Available at https://regcheck.app/.

DeBruine, L., & Lakens, D. (2025). *Papercheck: Check scientific papers for best practices* [Manual].

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Gohel, D., & Skintzos, P. (2025). *Flextable: Functions for tabular reporting* [Manual]. https://doi.org/10.32614/CRAN.package.flextable

Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

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

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23. https://doi.org/10.1145/3458754

Heathers, J. A., Anaya, J., Van Der Zee, T., & Brown, N. J. (2018). *Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE).* https://doi.org/10.7287/peerj.preprints.26968v1

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*(6), 1033–1037. https://doi.org/10.3758/BF03213921

Isager, P. M., & Fitzgerald, J. (2025). *Three-Sided Testing to Establish Practical Significance:*

A Tutorial. https://doi.org/10.31234/osf.io/8y925

Jeroen, O. (2025). *Magick: Advanced Graphics and Image-Processing in R* [Manual]. https://doi.org/10.32614/CRAN.package.magick

Lakens, D. (2021). The Practical Alternative to the *p* Value Is the Correctly Used *p* Value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

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

Murphy, S. L., Merz, R., Reimann, L.-E., & Fernández, A. (2025). Nonsignificance misinterpreted as an effect's absence in psychology: Prevalence and temporal analyses. *Royal Society Open Science*, *12*(3), 242167. https://doi.org/10.1098/rsos.242167

Nuijten, M. B., & Epskamp, S. (2024). *Statcheck: Extract statistics from articles and recompute p-values. R package version 1.5.0.* Web implementation at https://statcheck.io.

Python Software Foundation. (2025). *Python: A dynamic, open source programming language* [Manual]. Python Software Foundation.

Qiu, Y., & details., authors/contributors. of the included software. S. file A. for. (2024). *Showtext: Using fonts more easily in R graphs* [Manual]. https://doi.org/10.32614/CRAN.package.showtext

Schervish, M. J. (1996). *P* Values: What They are and What They are Not. *The American Statistician*, *50*(3), 203–206. https://doi.org/10.1080/00031305.1996.10474380

Stillman, D. (2019). *Retracted item notifications with Retraction Watch integration.* https://www.zotero.org/blog/retracted-item-notifications/.

Team, R. C. (2025). *R: A Language and Environment for Statistical Computing* [Manual]. R

Foundation for Statistical Computing.

van Abkoude, T. (2025). *Causal Confusion: How LLMs Can Improve Causal Language in Research Communication* [Master's {{Thesis}}].

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2025). *Readxl: Read excel files* [Manual]. https://doi.org/10.32614/CRAN.package.readxl

William Revelle. (2025). *Psych: Procedures for psychological, psychometric, and personality research* [Manual]. Northwestern University.

Table 2

Incorrect SciBERT Classifications

Model Prediction	Statement
False Negative	Although the sensitivity for the not-learned set was statistically comparable to the prelearning baseline, t(44) = 1.95, p = .162, d = 0.29, the learned set revealed significantly higher scores compared with both the not-learned set, t(44) = 2.56, p < .04, d = 0.38, and the prelearning set, t(44) = 4.51, p < .001, d = 0.67 (all comparisons performed with Bonferroni correction; see Fig. 3a).
False Negative	However, CS type did not interact significantly with the contrast between the uninformed and random groups; the uninformed group showed no better differentiation than the random group, F(1, 76) = 1.29, p = .26, 95% CI = [-0.16, 0.57].
False Negative	Gender, trait aggression, and endogenous testosterone did not affect these behavioral congruency effects on RTs and accuracy, and aggression and endogenous testosterone were not significantly correlated (r = .046, p = .45).
False Positive	The main effect of illness recency did not meet our preregistered threshold (p < .025)-recently ill: M = 661 ms, SD = 197; not recently ill: M = 626 ms, SD = 153, F(1, 400) = 4.23, Î· p 2 = .010, 90% CI = [.000, .039], p = .040-nor did the interaction between illness recency and face type (disfigured vs. typical), F(1, 400) = 1.87, Î· p 2 = .005, 90% CI = [.000, .027], p = .173.

Note. Examples for incorrect classifications of the trained SciBERT model on the test data.

Table 3

Examples of Incorrect LLM Corrections

Example	Statement Type	Statement	Label
Example 1 to 1	Original	However, when the target was uncommitted, participants no longer preferred the seeker, p = .849, d = -0.03, rating both targets equally.	1
	LLM-Corrected	However, when the target was uncommitted, participants showed no significant preference for the seeker over the other target, p = .849, d = -0.03, rating both targets equally.	1
Example 0 to 1	Original	By contrast, leader narcissism levels were not significantly related to follower-perceived leader authority, b = 0.04, 95% CI = [-0.04, 0.13], β = 0.10, p = .304, and effectiveness, b = 0.04, 95% CI = [-0.06, 0.14], β = 0.08, p = .398.	0
	LLM-Corrected	By contrast, leader narcissism levels were not significantly related to follower-perceived leader authority, b = 0.04, 95% CI = [-0.04, 0.13], Î² = 0.10, p = .304, and effectiveness, b = 0.04, 95% CI = [-0.06, 0.14], Î² = 0.08, p = .398, indicating that the relationship between leader narcissism and these outcomes may be negligible, but the data do not provide sufficient evidence to conclude that there is no relationship.	1

Note. Table of original and LLM-corrected statements that were classified as incorrect. In the examples '0' refers to correct and '1' to incorrect.

Table 4

Examples of Correct LLM Corrections

Example	Statement Type	Statement	Label
Example 1 to 0	Original	Only in the twin sample did a lower age of onset predict higher levels of melody discrimination (β = -0.09, p = .23 in musicians; β = -0.05, p = .001 in twins).	1
	LLM-Corrected	Only in the twin sample was there a significant association between a lower age of onset and higher levels of melody discrimination (β = -0.09, p = .23 in musicians; β = -0.05, p = .001 in twins), whereas in musicians, the association was not statistically significant.	0
Example 0 to 0	Original	The years-of-experience single effect was not significant (β = 0.001, 95% CI = [-0.07, 0.07], p = .976).	0
	LLM-Corrected	The years-of-experience single effect was not significant (β = 0.001, 95% CI = [-0.07, 0.07], p = .976), indicating that the data do not provide sufficient evidence to reject the null hypothesis of no effect.	0

Note. Table of original and LLM-corrected statements that were classified as incorrect. In the examples '0' refers to correct and '1' to incorrect.