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 an automated approach: 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 models performed well, with the standard BERT model achieving the highest macro F1 score of .92. Finally, the optional correction step proved effective in a validation set of 80 incorrect and 20 correcct statements: 93 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 (Aczel et al., 2018; Hoekstra et al., 2006; Murphy et al., 2025). So, how might we be able to finally overcome them?

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 misinterpretations. I focus on the misinterpretation of statistically nonsignificant results as the absence of an effect because it arguably has the strongest impact on researchers' conclusions and is the most extensively studied misinterpretation of *p* values (Lakens, 2021). Similarly, my previous work has developed clear criteria for classifying this misinterpretation (Murphy et al., 2025). I demonstrate how this automated approach may help us to finally overcome this misinterpretation.

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 (e.g., Greenland, 2019; 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).

In this study I zoom in on one specific misinterpretation: concluding *no effect* based on a statistically nonsignificant finding. Many studies have previously shown that this misinterpretation remains highly prevalent across time and sub-domains of psychology (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). These findings highlight that the situation seems not to have greatly improved despite continuous calls to reflect on statistical interpretations of nonsignificant results (e.g., Altman & Bland, 1995; Gelman & Stern, 2006) and and increasing advocacy for alternative analytical approaches that enable researchers to make informed claims about effects being practically equivalent to zero (e.g., Dienes, 2014; Lakens et al., 2018).

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 persistent prevalence of the misinterpretation examined in this study indicates that calls for improved education alone have not been sufficient to address the issue (Murphy et al., 2025). This is complemented by research showing that many misinterpretations of *p* values are shared among psychology students and teachers (Badenes-Ribera et al., 2016; Haller & Krauss, 2002). Recognizing the limitations of education alone, 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 proposed solutions 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 contexts we encounter systems like this frequently, be it automatic braking systems in cars, word processors that flag spelling and grammar mistakes, or email clients that filter out malware and phishing attempts. Analogously, 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. As AI continues to develop, we can expect these types of automated solutions to become increasingly sophisticated and common.

Following this trend, DeBruine and Lakens (2025) developed Papercheck, an R package, which allows users to run a battery of automated 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.

–> I originally had a short summary paragraph here to, again, repeat the “research question”/aim and to transition to the methods more smoothly, but Daniel (my supervisor) said it wasn’t necessary and too much repetition. Let me know if you feel like its missing here! <–

2 Methods

2.1 The Thee-Step Pipeline

To provide context for the data used in this study, I first outline the three sub-steps of the proposed pipeline. Statements from scientific articles need to be reliably detected, classified, and finally, if desired, corrected. For each step, I applied specific methods that were best suited to achieve the respective goal.

To detect statements I used rule-based regular expressions (RegEx) and searched articles' results sections to detect these expressions. Effectively, RegEx searchers are advanced Ctrl+F searches, where a user can 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 has a module that detects almost all *p* values (see Section 3.1 for examples currently not detected) based on RegEx searches. Using this module, I created a subset of all *p* values equal to or above .05.[1] I then expanded the extracted nonsignificant *p* values to the full sentence and added +/- one sentence as context in case of extraction errors (incomplete statements).

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 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 is an even more specific pretrained language model, having been trained exclusively on biomedical abstracts and full-text articles from the PubMed database (Gu et al.,

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

2022). These models were trained and evaluated on their ability to distinguish between correct and incorrect interpretations of nonsignificant results in scientific writing. The models' hyperparameters (e.g., learning rate, batch size) were informed by established defaults in the field (see https://huggingface.co/docs/transformers/en/main_classes/trainer) and relevant tutorials (More, 2025; Talebi, 2024), with further refinements to improve the models' prediction performance.

Lastly, in the application of this framework, statements classified as incorrect by the best-performing BERT model would be sent to a large language model (LLM) for correction. However, to assess how the LLM handles both genuinely incorrect statements and those misclassified as incorrect automatically, I submitted both correct and incorrect statements coded by me to the LLM in this study (see Section 2.2 for details).

2.2 Validation Process and Performance Metrics

To assess the effectiveness of each automated approach, I compared their outputs to human-coded ground truth and calculated appropriate reliability and performance metrics. The validation process was conducted separately for statement detection, classification, and correction.

2.2.1 Statement Detection

Firstly, to ensure that the statement detection process caught all statements with nonsignificant *p* values in articles' results sections, I manually extracted these statements from 25 (10%; randomly chosen) 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). I then coded whether statements I found were also extracted with the automated RegEx search.

2.2.2 Statement Classification

For the training of the BERT models and to assess their final performance, I labeled all automatically extracted statements that were detected from an article's results section from Papercheck's sample library. This resulted in 960 statements in total. Of these, 419 were classified as containing a correct *p* value interpretation by me, and 353 were classified as incorrect *p* value misinterpretations. 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 falsely flagged as containing a nonsignificant *p* value (e.g., significant *p* values or generic '*p* > .05 indicated by symbol xy' statements from table/figure notes; 19 in total), or due to a combination of these or other reasons (46).

Before actually training a model, the labeled data was split into three parts: a test set (20%) used for the final evaluation of the model, a training set (72%, or 90% of the remaining 80%) that the model used 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) to prevent overfitting to the training data. The number of correct and incorrect interpretations was balanced in each of these parts (originally there were more correct than incorrect interpretations) to ensure that the model would not overfit to this class-imbalance (and only predict the majority class).

Before training the model, the labeled data was split into three parts: a test set (20%) for final model evaluation, a training set (72%, or 90% of the remaining 80%) for learning and parameter adjustment, and a validation set (8%, or 10% of the 80%) for calculating evaluation metrics after each epoch (i.e., one full cycle of the model processing the training data) and monitoring overfitting. To address the original class imbalance (with more correct than incorrect statements), the number of correct and incorrect interpretations was balanced in each set, ensuring the model would not simply learn to predict the majority class.

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 to prevent the model from overfitting to the training data. 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 consecutive epochs. Ultimately, the longest number of training epochs was 7. For the 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 (incorrect and correct). To summarize overall performance across the two classes, I calculated the unweighted average of the two F1 scores (macro-F1 score).

2.2.2 Statement Correction

Lastly, I reviewed 100 statements that were sent to an LLM for correction to evaluate whether the revised statements were correct. Of these, 80 had previously been labeled incorrect and 20 correct, allowing me to examine how the LLM handled false positives from the automated classification. To communicate with the models, I used Papercheck, which relies on the Groq API (available at https://groq.com/). I tested two LLMs - 'llama-3.3-70b-versatile' (created 03-09-2023) and 'openai/gpt-oss-120b' (created 05-08-2025) - and applied two prompts to the full validation dataset of 100 statements. This resulted in three iterations: (1) the initial prompt with the 'llama-3.3-70b-versatile' model, (2) the same prompt with 'openai/gpt-oss-120b', and (3) a refined prompt, developed through preliminary tests on subsets of the 100 statements, with 'openai/gpt-oss-120b'.

Both prompt versions first explained to the LLM that it would receive a statement containing at least one misinterpretation of a nonsignificant finding as the absence of an effect, and instructed it to revise only the part of the statement containing this misinterpretation, leaving the rest unchanged. The refined prompt included additional guidance on phrasing to avoid, based on common errors that persisted in revisions from the initial prompt. Finally, the initial prompt instructed the LLM to respond with "NO CORRECTION POSSIBLE" if it found no nonsignificant *p* value or interpretation of it to account for possible errors during the automatic detection of statements. This instruction was removed in the refined prompt because the LLM overused this option, which blurred the distinction between statement detection, classification, and correction. Both prompts are available on GitHub (see Section 2.3).

2.3 Software

All scripts for this study 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 & Lakens, 2025) for accessing the 250 open access articles, preprocess them and for communication with the LLMs, *readxl*

(Version 1.4.5; Wickham & 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 & Skintzos, 2025), *magick* (Version 2.8.7; Jeroen, 2025), *papaja* (Version 0.1.3; Aust & Barth, 2024) and *showtext* (Version 0.9-7; Qiu & for details., 2024) to create APA-formatted tables and figures.

All scripts and data to reproduce and use the trained BERT models (Python), analyse the results and validity checks (R and Python), recreate this manuscript (Quarto Markdown in R Studio with the apaquarto extension available at: https://wjschne.github.io/apaquarto/), as well as the list of Python libraries used to train the BERT models are available in this GitHub repository, together with instructions on how to set it up: LINK.

Due to the project's iterative nature and since no inferential statistical tests were performed, this study was not preregistered. The original project proposal can also be found in the GitHub repository.

3 Results

3.1 Detection Accuracy

By manually reviewing 25 articles from Papercheck's sample library, I identified 179 statements containing a nonsignificant *p* value. The automated RegEx search fully detected 130 (73 %) of these, and partially detected 6 due to extraction errors, often caused by PDF formatting (page breaks, figures, or footnotes). The search also produced 3 false positives - statements incorrectly labeled as coming from the results section, but actually originating from other sections, or from table or figure notes. Note, however, that most of the 49 (partially) missed statements were due to specific ways of writing (or not writing) the *p* value: 31 were missed because the *p* value was written as 'p_s', and 8 were missed because the authors wrote 'n.s.' instead of the nonsignificant *p* value. Excluding these two types of reporting, the overall agreement between the automated and manual approaches would have been 93 %.

Lastly, the remaining 10 missed statements were due to pdf formatting issues such as figures, tables, footnotes, page breaks, or unusual characters within the statement that

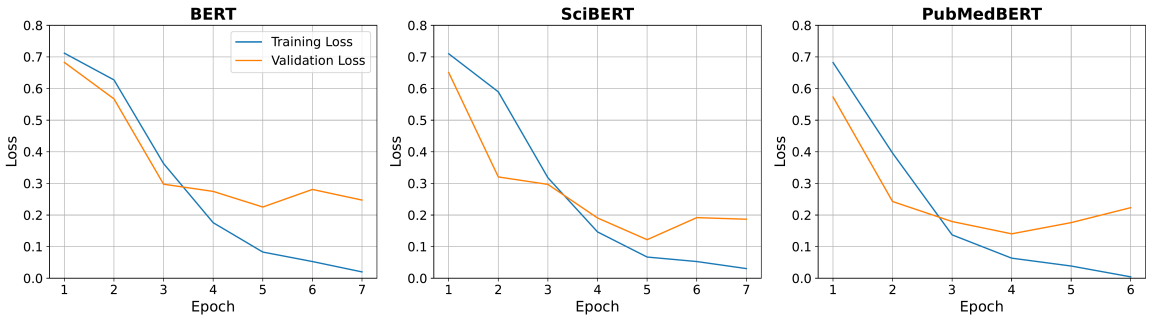
interfered with the statement detection (9 in total).[2]

3.2 Classification Performance

Figure 1 shows the training and validation loss curves for the three BERT models across their training epochs. The standard BERT model was trained for a total of 7 epochs before early stopping was triggered due to a lack of improvement in validation loss for two consecutive epochs. The model from epoch 5 was therefore selected as the best-performing one. Similarly, SciBERT and PubMedBERT reached the lowest validation loss after epochs 5 and 4, respectively. As shown in the figure, the training loss consistently decreased throughout training for all three models, as expected since the models were optimized to fit the training data. In contrast, the validation loss plateaued in all models before rising again, indicating that further improvements on the training data no longer translated to better performance on the unseen validation data and may signal the onset of overfitting.

Figure 1

Training and Validation Loss Curves



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

Moving to model performance, Figure 2 displays the number of correctly and incorrectly classified statements for each model. Overall, all three models performed well, with the standard BERT model showing the fewest misclassifications (2 false positives and 9

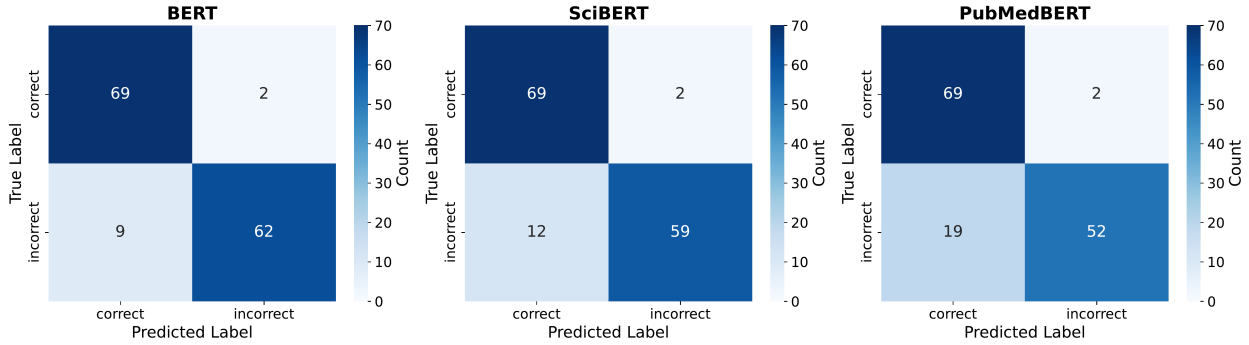
[2] I could not find one statement that was detected automatically in the article's pdf. My current theory is that this was an artifact from when the pdf was compiled and might be from a different article even, once again highlighting how impractical the pdf format is in times of increasing automation.

false negatives), while SciBERT and PubMedBERT made 14 and 21 false classifications in total, respectively. All models also tended to predict more statements as correct than incorrect: for example, the standard BERT model classified 78 statements as correct and 64 as incorrect, with this difference being even more pronounced in SciBERT (81 vs. 61) and especially PubMedBERT (88 vs. 54).

These results are further summarized in Table 1. The standard BERT model achieved the highest macro-F1 score (.92), with SciBERT and PubMedBERT scoring slightly lower (.90 and .85). Across all models, performance was stronger for predicting correct statements than incorrect ones, as indicated by higher F1 scores and recall in the 'correct' class. This pattern, again, underlines that the models tended to overidentify statements as correct rather than incorrect.

Figure 2

Confusion Matrices



Note. Confusion matrices of the three trained BERT models. Overall, the standard BERT model performed best with a macro-F1 score of .92.

Table 2 shows statements misclassified by all three models to illustrate common sources of difficulty. Potential causes of these misclassifications will be explored in the discussion.

3.3 Correction Evaluation

Of the 100 statements that the LLM was instructed to correct 85 were evaluated as correct in the first iteration (initial prompt and older 'llama-3.3-70b-versatile' model). Notably, 2 of the 20 already correct statements were turned incorrect by the LLM, and 13 of the 80 incorrect statements remained incorrect. Using the newer 'openai/gpt-oss-120b' model,

Table 1

Model Performance

	BERT			SciBERT			PubMedBERT		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
Correct Class	.88	.97	.93	.85	.97	.91	.78	.97	.87
Incorrect Class	.97	.87	.92	.97	.83	.89	.96	.73	.83
Macro F1 score			.92			.90			.85

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

79 statements were evaluated as correct after LLM-revision. However, the model overused the 'NO CORRECTION POSSIBLE' option (originally intended to resolve possible detection errors) which resulted in 8 correct and 2 incorrect statements being left unrevised. In the final iteration, with the newer model and revised prompt, 93 of the 100 LLM-revised statements were correct. Of the remaining 7 statements, four (one originally correct and three incorrect) described the results as 'compatible with effects of exactly or around zero' - an interpretation that may be technically defensible, but still fails to adequately acknowledge the uncertainty in the test result. In one originally correct statement, the LLM removed all *p* values and described the effect as if they were significant. The remaining two were originally incorrect statements, one of which simply remained incorrect, while the other one was altered to the point that it no longer conveyed the original meaning (originally referring to model fit, which was not clear in the revision). Examples of both poor and strong LLM-revisions from this final iteration are shown 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 detecting, classifying and correcting misinterpretations of nonsignificant results as evidence for the absence of an effect. The pipeline combines (1) rule-based RegEx searches to identify candidate statements, (2) fine-tuned BERT models to classify them as correct or incorrect, and (3) LLMs to generate revised phrasings. Although clear areas for improvement emerged, each component performed well overall. This demonstrates the potential of hybrid, step-wise

Moving to the sub-steps, the RegEx-based statement detection phase demonstrated that simple, rule-based searches can effectively flag a large proportion of candidate interpretations. Although formatting issues in pdfs made the correct extraction of these misinterpretations impossible in some cases, the vast majority of statements were automatically detected. In addition, this study revealed straightforward issues in the current approach (e.g., as "p = n.s." or with subscripted 'p_s') that can be fixed with minimal adjustments, further enhancing the detection accuracy.

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. Interestingly, the standard BERT model showed the best overall performance, potentially reflecting that SciBERT's and PubMedBERT's more domain-specific knowledge might not have been necessary for the classification of these statements.

Finally, the correction step demonstrated that LLMs can propose useful alternative phrasings to misinterpretations. While prompt engineering will be necessary to reduce occasional mistakes (especially those that turning correct statements into incorrect ones), this step illustrates the potential of integrating generative AI into targeted scientific writing support. Whether such corrections are genuinely valued by authors, however, remains an open question, needing further discussion.

4.2 Limitations and Challenges

A key limitation is that the pipeline was tested step by step rather than as a fully integrated system. While detection, classification, and correction each worked well independently, cascading errors are inevitable once these steps are combined. As a result, the overall accuracy in end-to-end use will likely be lower than suggested by isolated evaluations. Nonetheless, any detected misinterpretation should alert authors that their interpretation, some possibly left unflagged, may require reconsideration.

Similarly, the RegEx-based approach proved efficient but also exposed the fragility of

rule-based methods when faced with variation in formatting. Cases such as 'p = n.s.' or p_s were frequently missed. These errors are relatively easy to address through refinements of the pattern set but highlight that rule-based detection will always struggle with rare or novel notations.

At the next step, the classification step performed strongly overall, but the use of a single annotator introduces subjectivity into the training labels. While I made efforts to standardize labels - often consulting a statistics expert (my supervisor) on difficult or borderline cases - ultimately, the classifications reflect my interpretation of what constitutes a misinterpretation. Ideally, multiple annotators and inter-rater agreement metrics would strengthen the reliability and generalizability of the training dataset. However, the fact that the fine-tuned BERT models generalized well to unseen data suggests that the labeling was systematic enough for the models to learn.

The standard BERT model consistently outperformed the more domain-specific SciBERT and PubMedBERT models, suggesting that this specialization might be hindering in this context. This difference could also be explained by the size of the models' training datasets: BERT was trained on 3.3 billion words (Devlin et al., 2019), whereas SciBERT and PubMedBERT were trained on 3.17 billion tokens (words or word fragments; Beltagy et al., 2019) and 3.1 billion words (Gu et al., 2022), respectively. BERT may therefore have achieved better results simply because its broader and larger training data allowed it to represent sentence structure more effectively. Of course, the difference could also stem from numerous other factors in the models' training configurations. Future research should examine whether this performance gap persists across related tasks.

Similarly, with respect to this study's training data (articles from Psychological Science published between 2013 and 2024), the findings should not be overgeneralized. While the trained BERT classifiers will likely detect misinterpretations in articles from other journals and time periods as well, the training dataset would nonetheless benefit from being expanded and diversified to cover a broader range of psychological research.

While the standard BERT model performed well overall, it still misclassified edge cases, particularly statements containing both correct and incorrect elements. For example,

'the difference was not significant, suggesting there is no effect' combines a factual result with a problematic inference. As can be seen in Table 2, all three models often classified such statements wholly as correct. Similarly, if correct statement did not use the word 'significant' explicitly, models sometimes misclassified these as incorrect. This illustrates the difficulty of reducing nuanced writing to a binary label. Splitting statements into parts or using non-binary classification approaches might therefore be useful.

A more practical challenge involves managing models' tradeoff between false positives and false negatives predictions. The current models aim to balance both for optimal macro performance (as seen in the results, this was not perfectly possible). 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 for predicting one label or the other. A future Papercheck module based on this work could incorporate such functionality to fit users' specific goals.

Lastly, the correction stage revealed both opportunities and challenges. LLMs were able to propose helpful revisions, but, as shown in Table 3, they also turned already-correct statements incorrect or produced overly generic revisions. Providing them with more context (e.g., +/- one sentence or the full paragraph) might improve their ability to generate accurate and nuanced corrections.

However, there is also the flip side to the aforementioned human factors perspective: should authors rely on AI-generated corrections at all? Unlike grammar mistakes that word processors automatically fox for us, misinterpretations of statistical results often stem from deeper conceptual misunderstandings. Automated corrections might inadvertently encourage passivity, shifting responsibility away from researchers' own critical engagement. In this sense, corrections should, if included at all, be seen as optional guidance, not authoritative replacements for careful reasoning.

4.3 Implications and Future Directions

The pipeline described in this study will be integrated into a new Papercheck module for identifying potential misinterpretations of nonsignificant results. Some clear aspects to improve have been detected in this study: Firstly, the current RegEx searches of Papercheck's 'all_p_values' module might not be optimized to detect all different ways in which a *p* values can be written. For example, the previously mentioned p_s is often used to refer to the smallest *p* value in some collection of tests. This is an example of usually irrelevant RegEx's that I will add to improve this automatic detection of candidate statements. Additionally, the dataset used to train the BERT models will also be expanded and re-checked by independent coders to ensure that the aspects the models do pick up are generalizable. Lastly, mistakes from the correction validity check of statements' LLM-revised corrections will be closely analyzed to inform further prompt engineering to reduce any mistakes.

A broader implication concerns how nonsignificant results should be reported in general. These interpretations are not free-form expressions that authors can adapt to suit their way of writing but rather formalized claims where certain phrasings are demonstrably misleading. Automated systems can support a shift toward clearer and more standardized reporting practices, but ultimately, researchers must recognize their own responsibility in this regard.

This issue is closely connected to the broader question of why such misinterpretations, despite decades of critique, remain highly prevalent, a phenomenon likely driven by multiple factors. Educational gaps leave many students and researchers uncertain about how to interpret *p* values correctly, as instructors may share the same misconceptions (Haller & Krauss, 2002) or textbooks may themselves misrepresent key statistical concepts (Cassidy et al., 2019). Similarly, researchers might have different philosophies of science and might disagree about how to interpret key statistics like the *p* value (Lakens, 2021). Finally, the widespread prevalence of misinterpretations itself creates a self-reinforcing cycle, with researchers adopting the language they encounter in published articles and thereby perpetuating the problem. Automated feedback systems cannot resolve these deeper causes, but they may, for now, assist individual authors by highlighting such mistakes in their own work.

Additionally, the pipeline's step-wise structure makes it easy to adapt to other classification or correction tasks than the one presented in this study. For instance, users could train custom classifiers to detect different issues in reporting practices (see van Abkoude, 2025 for an application to problematic causal language). In practice, this would involve specifying RegEx patterns that capture the target aspects, training classifiers to label them as correct or incorrect, and, if desired, creating a prompt to generate corrections. Depending on the issue at hand, such classifiers could also be trained on existing hand-labeled datasets from meta-scientific studies where researchers coded specific practices or mistakes (e.g., Aczel et al., 2018).

Finally, the most important next step for this project will be conducting qualitative user studies to explore how authors would prefer a tool this to be designed and implemented. A central question will, again, be the role of the optional correction feature - whether authors find value in receiving suggested corrections or whether simple flagging is sufficient. These studies could also reveal where customization is most useful (e.g., varying levels of strictness, setting a personal alpha level instead of the conventional 5%, etc.). In addition, experimental evaluations would help assess whether the tool reduces the prevalence of misinterpretations and increases authors' awareness of them.

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 and conduct qualitative user studies to inform the tool's implementation. With further refinements, automated manuscript checks like the one presented in this article could substantially improve the accuracy of scientific reporting, while also providing a valuable resource for large-scale meta-scientific analyses.

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 very 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 to 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

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485–485. https://doi.org/10.1136/bmj.311.7003.485

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

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01247

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

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly. *Advances in Methods and Practices in Psychological*

Science, *2*(3), 233–239. https://doi.org/10.1177/2515245919858072

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

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

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

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

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. (2019). Valid *P* -Values Behave Exactly as They Should: Some Misleading Criticisms of *P* -Values and Their Resolution With *S* -Values. *The American Statistician*, *73*(sup1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

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

Haller, H., & Krauss, S. (2002). *Misinterpretations of Significance: A problem students share with their teachers?* https://doi.org/10.5283/EPUB.34338

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

More, R. (2025). *Fine-Tuning BERT for Text Classification Using Hugging Face Transformers*.

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., & for details., authors/contributors. of the included software. S. file A. (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*.

Talebi, S. (2024). *Fine-Tuning BERT for Text Classification*.

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 Model 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 statements that were incorrectly classified by all three BERT models.

Table 3

Examples of Incorrect LLM-Revisions

Labels	Statement Type	Statement	Label
	Original	Although the PTSD group showed a significant facilitation effect (M = 5.9%), F(1, 15) = 7.525, p = .015, Î· p 2 = .334, and the control group did not (M = 2.8%, n.s.), there was no significant group difference in facilitation, F(1, 30) = 1.215, p = .279, Î· p 2 = .039 (Fig. 3a).	0
From 0 to 0	LLM-revised	Although the PTSD group showed a significant facilitation effect (M = 5.9 %), F(1, 15) = 7.525, p = .015, ηp² = .334, the control group did not reach statistical significance (M = 2.8 %). The group difference in facilitation was not statistically significant, F(1, 30) = 1.215, p = .279, ηp² = .039 (Fig. 3a).	0
	Original	Cortisol concentrations were comparable at baseline, t(55) = 1.01, p = .317, d = 0.27, but significantly elevated in the stress group relative to the control group both 30 min after the treatment, t(55) = 3.55, p corr = .003, d = 0.94, and 45 min after the treatment, t(55) = 2.52, p corr = .045, d = 0.67 (Table 1).	1
From 1 to 0	LLM-revised	Cortisol concentrations were not significantly different at baseline, t(55) = 1.01, p = .317, d = 0.27, but were significantly elevated in the stress group relative to the control group both 30 min after the treatment, t(55) = 3.55, p corr = .003, d = 0.94, and 45 min after the treatment, t(55) = 2.52, p corr = .045, d = 0.67 (Table 1).	0

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

Table 4

Examples of Correct LLM-Revisions

Labels	Statement Type	Statement	Label
	Original	Non-attraction-effect choices were very rare and, like decoy choices, they likely reflect noisy responding; they were more common among people with lower cognitive ability, b = 0.689, 95% CI = [0.357, 1.021], p < .001, but did not differ between the ASC and control groups, b = -0.037, 95% CI = [-0.381, 0.307], p = .834, and were unrelated to age, b = 0.014, 95% CI = [-0.295, 0.324], p = .929, or gender, b = -0.229, 95% CI = [-0.528, 0.069], p = .132.	1
	LLM-revised	Non-attraction-effect choices were very rare and, like decoy choices, they likely reflect noisy responding; they were more common among people with lower cognitive ability (b = 0.689, 95 % CI = [0.357, 1.021], p < .001). The estimated difference between the ASC and control groups was b = -0.037 (95 % CI = [-0.381, 0.307], p = .834). The estimated association with age was b = 0.014 (95 % CI = [-0.295, 0.324], p = .929), and the estimated association with gender was b = -0.229 (95 % CI = [-0.528, 0.069], p = .132).	0
From 1 to 0	Original	This preference, however, disappeared in the experimental condition, F(1, 170) = 2.32, p = .130, Î· p 2 = .01.	1
	LLM-revised	In the experimental condition, the preference was not statistically significant, F(1, 170) = 2.32, p = .130, ηp^2 = .01.	0

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