**Notes on the labeling/coding**

* Maybe exclude all statements that are not from results section?
  + Still include those from discussion (even though they have p values in them?)
  + Some statements are actually okay but are flagged as being from a figure not or sth. 🡪 Check of simply exclude!
* Difference between strict and kind label
  + Strict: anything that might lead readers to believe that the effect is null; similar to labels in Murphy et al. (2025) and Aczel et al (2018)
    - May include negations in both test descriptions and test result interpretations
  + Kind: only direct and explicit test result interpretations as no effect. Negations of test descriptions are okay (e.g., there was no main effect/interaction??)
    - Check again with Daniel, bc it feels strange to code this as correct
    - Clarify what he means
* ~~“Ã¢â€°Â¤” etc. seem to be errors. I think it should men < or >. Need to check!~~
* For now, I only coded statements that had “results” as their section! Might want to check the others later!
* I did not include statements where you needed more context than the 1 sentence (e.g., something like “This was also true for XY (p = .524).”) bc I only wanted to include one sentence statements for now
* Statements like “X had a significant effect, but Y did not” sound incorrect to me, but I will label them as correct bc of the “significant”
* “No evidence” statements are strange; absence of evidence is okay, but reading it still sounds like “seems like there’s nothing there”
* If a statement contains both a nonsig. p value and a BF, do I even code them as incorrect?
* How to deal with statements that mention equivalence tests?
* Problem that some statements contained both correct and incorrect interpretations 🡪 might be tricky for the classifier
* Marginally significant key word if they interpret nonsig effects as significant/real effects (not only when they say ‘trend’ or ‘marginally’!)
* ‘We found no effect…’ coded as correct
  + I kinda wanna code this as incorrect now…
* Idea: Instead of a classifier, maybe just use something like distinct\_words() but something that identifies the most common word pairs (e.g., ‘not significant’, ‘no effect’, ‘not predict’, …)! Use this for automated coding and see how much it aligns with the hand-coded examples
* Statements like ‘We observed no such increase…’ maybe wouldn’t be 100% incorrect, but I would still like them to be coded as 1 so that they are highlighted and written in a more correct way with the automated approach
* Very subtle difference between “we did not find X” (correct) and “we found no difference” (incorrect) might be really hard for a model to pick up
  + Correction: Maybe not always incorrect?
  + See: “*Here, we found no main effect of experiment, F(1, 78) = 0.18, p = 0.67, Î· p 2 < .01), a main effect of sound, F(1, 78) = 24.04, p < .0001, Î· p 2 = .24, and a significant interaction, F(1, 78) = 4.21, p = .0435, Î· p 2 = .05.”*
  + Really tricky…
* Tricky balance for some statements like “No effect was found” 🡪 Code as correct bc it doesn’t strongly hint towards the idea that there is no effect OR code as incorrect, bc it should still be interpreted differently ideally?
* When the LLM generates a new (and ideally correct) version of a statement, it should also check if the nonsig p value in the statement seems to even be interpreted. If not, it should say something like “You don’t seem to interpret the nonsignificant result in this statement. Why not?” or so
* Again about the LLM correction of mistakes: I first thought that it might be good if the LLM takes all the context of the statement (e.g., the full paragraph) into account. But then it would need to rewrite the full paragraph, or that would probably create the best result. Since I’m not sure that that is how LLM should be used, for now, I only want them to correct the mistake within the sentence. Researchers can then correct the paragraph themselves.
* The more statements I read, the more I think this is useless. Something like “the test showed that there was no significant correlation” sounds so wrong to me, and it should be something like “the difference that is clearly there in the sample did not reach stat significance, we therefore cant rule out that the effect might be zero” or so…
* ‘we found no reliable effect/difference’ will be coded as correct (similar to ‘significant’)
* If they just don’t write ‘significantly’ in the second part of a sentence, its still coded as correct: e.g., “The effect of X was significant, but not for Y/but Y was not” etc.
* ‘Analyses revealed that there was no effect’ would be incorrect, whereas ‘Analyses revealed no effect’ would be correct
* What if researchers use equivalence tests?
  + Still use classifier that might highlight something like “no association”, similar to how Word also underlines some strange words it doesn’t know for example
* I coded statements as ‘-99’ if you would need more context than just the one sentence (simplest classifier for now)
* As an idea: run classifier on Marphy et al. (2025) statements and calculate agreement
* Limitation: I cant know whether some of the researchers might have also done equivalence testing, but since the classifier should be a help for writing assistance, researchers can just dismiss any ‘false flag’ and keep their manuscript as is (or incorporate the equivalence test right into the sentence to avoid confusion)!
* What might confuse the model are statements with two interpretations, one of which is correct and the other incorrect
* - Idea: Check whether articles that cite an article with a null finding (and cite the null finding) also misinterpret this null finding (or copy the misinterpretations of the original study)
  + tricky to implement and probably out of the scope of this project!
* No effect = 1; 'at chance level' = 1
* Discussion
* Did not reveal is okay; revealed no effect is not

**Notes from manually extracting statements (detection\_check):**

* Miss some statements when there’s no p value, but e.g., a confidence interval that includes the 0
* Misses any ‘n.s.’ instead of p value (e.g., ‘Participants who were trained in face perception showed significantly improved performance in face perception at Test Session 2, z = 3.92, p < .003, but not at Test Session 3, z = 2.79, n.s.’)
* There are also statements without any p value (e.g., ‘None of the contrasts for face perception training was significant (see Table S3 in the Supplemental Material).’)
* One statement broke bc of a page break inside the statement; note that papercheck usually handels these well!
* Some studies just report all p values (e.g., in large regression models) in a table an then only cite the estimate name or sth (NOT the p value), but still interpret significance; not depicted in the current approach
* Feels more like ‘either papercheck can get statements in your article, or it can’t’ which might be a good argument. 🡪 Maybe include warning if it cant find any nonsig. p values? 🡪 “We found no nonsig. p values in your manuscript. This might be an error. Best, you check yourself.”

**Notes from manually checking LLM correction:**

* LLM can only fix so much: sometimes there are other misinterpretations in a statement, e.g., that a sig and nonsig finding are conflicting. LLM usually doesn’t fix that!
* It’s really interesting to see what the LLM does with already correct statements. It adds in part elaborate explanations what the nonsig p value could mean (e.g., power issue, data variability, etc.
* I don’t yet assess how well the LLM deals with -99 statements!
* LLM often just fixes a ‘no effect’ to ‘no significant effect’ and ignores any ‘, suggesting …’ bits
* With the instructions to change as little as possible, some statements are not fixable, because they 100% aim at a ‘there’s nothing there’ interpretation.
* LLM doesn't like to change 'similarly' etc.