**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
* 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.