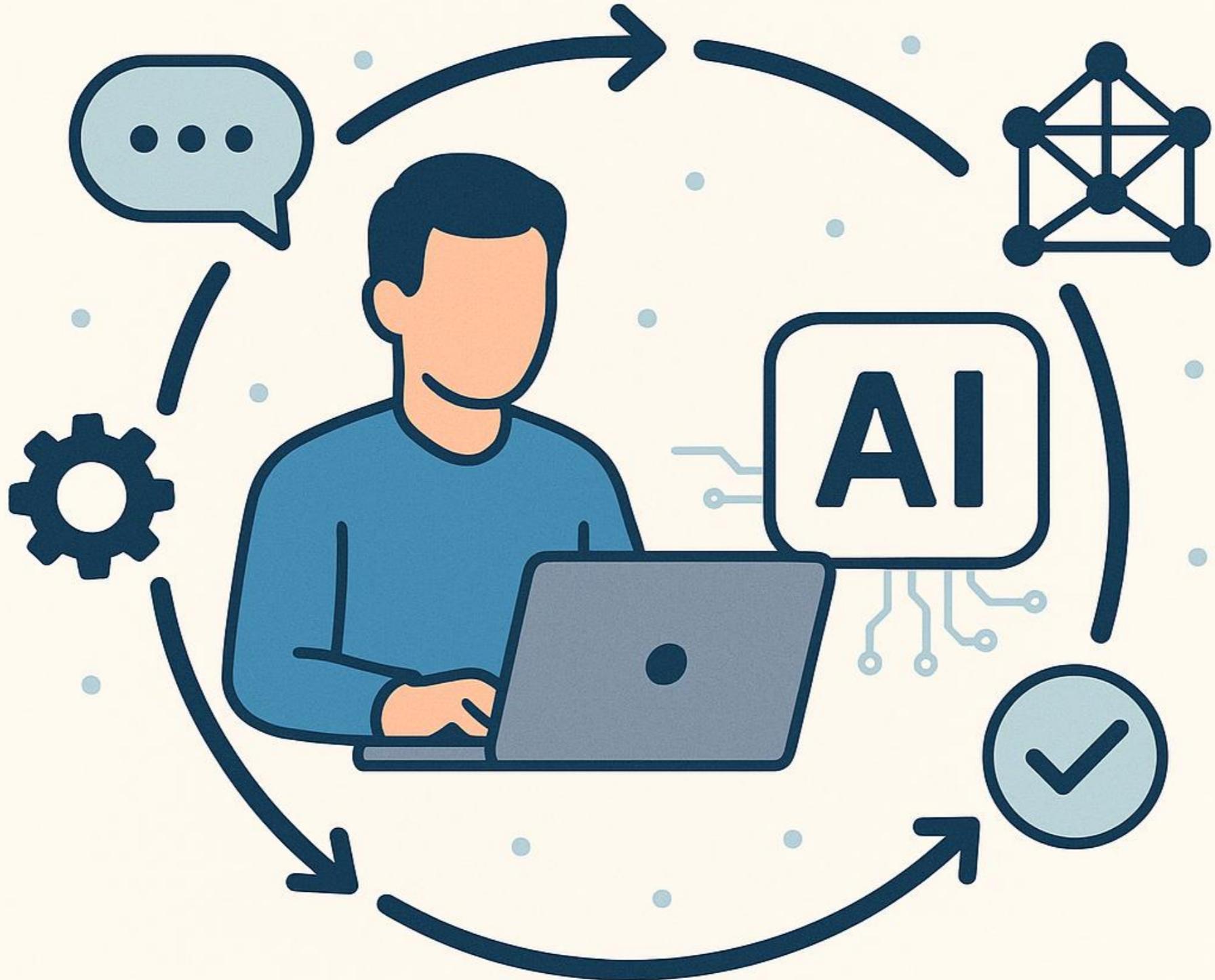


GenAI VOOR BACHELOR- EN MASTERPROEVEN: DATA-ANALYSE

Uitwisselsessie met diverse experten – 4 december 2025

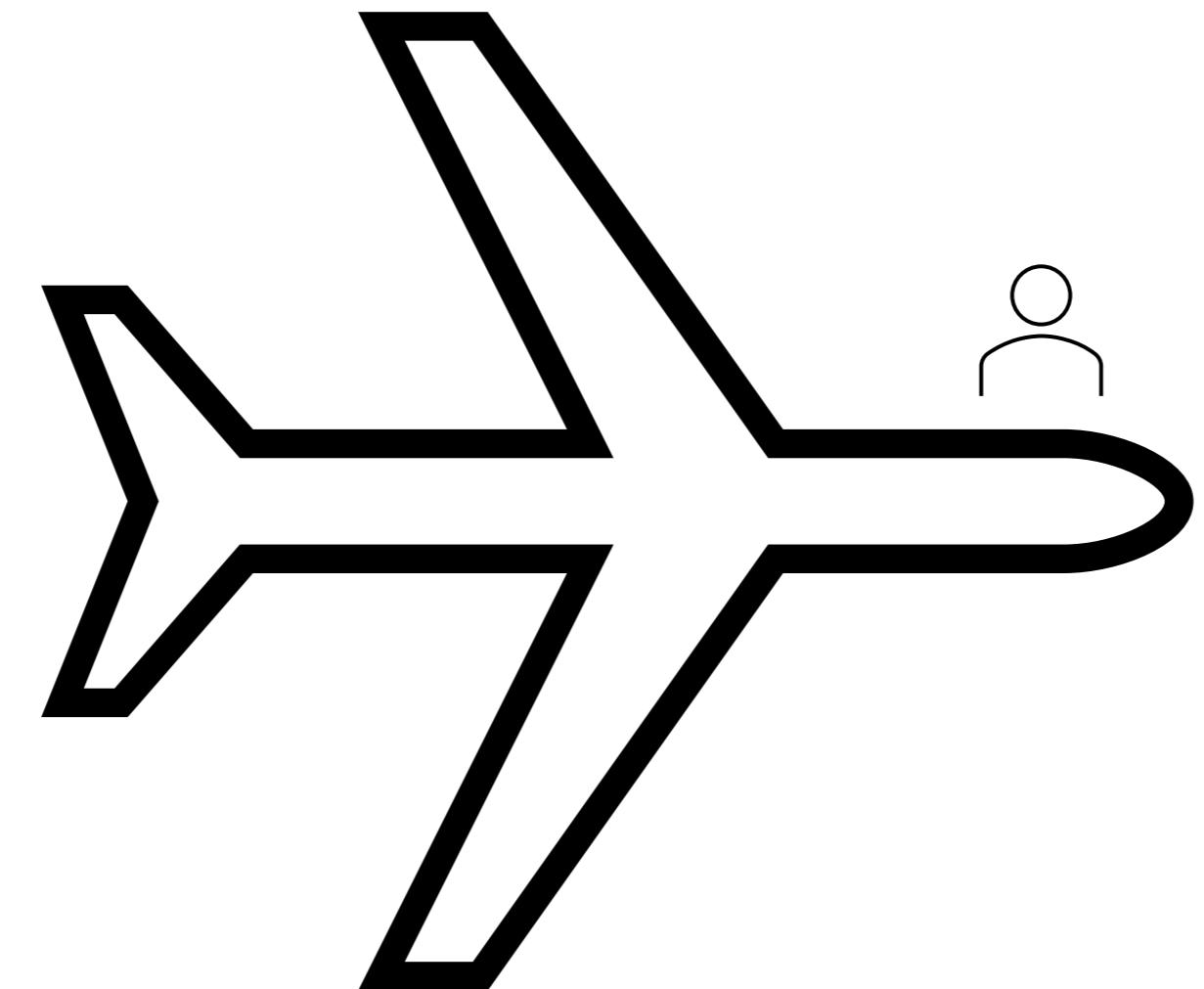
Ondersteuningsaanbod GenAI voor lesgevers en studenten LW – AJ25-26, 1ste semester

DISCLAIMER



HUMAN IN THE LOOP

HUMAN IN THE COCKPIT



COGNITIVE COST OF BLIND AI USE

As the educational impact of LLM use only begins to settle with the general population, in this preliminary study we demonstrate the pressing matter to explore further any potential changes in learning skills based on the results of our study. The use of LLM had a measurable impact on our participants, and while the benefits were initially apparent, as we demonstrated over the course of 4 sessions, which took place over 4 months, the LLM group's participants performed worse than their counterparts in the Brain-only group at all levels: neural, linguistic, scoring.

Taco Bell rethinks AI drive-through after man orders 18,000 waters

29 August 2025

Share  Save 

Shiona McCallum Senior Tech Reporter



Getty Images

Meta investigated over AI having 'sensual' chats with children

18 August 2025

Share  Save 

Charlotte Edwards Technology reporter, BBC News



Getty Images

ADDED VALUE OF AI “AUTOPilot”?

- Starting point: is use of AI necessary / beneficial?
- “Don’t ask ChatGPT how much $2+2$ is” mindset
- Creating graphic in Excel versus creating graphic with AI-driven tool
- Talking to supervisor versus talking to general-purpose chatbot

OVERVIEW

MENU

1. Starters

- Guidelines on GenAI
- Warm-up case

2. Main course: cases presented by experts

3. Dessert: cases submitted by you

SNACKS ON THE PANEL

- Romeo De Timmerman
- Koen Plevoets
- Tess Dejaeghere

1. STARTERS

GUIDELINES ON GENERATIVE AI

GHENT UNIVERSITY

- Overarching guidelines → “responsible” use is allowed
 - University: <https://www.ugent.be/genai>
 - Faculty: Ufora “[INFOSITE Faculteit Letteren en Wijsbegeerte](#)”
- ***Leerpad for teachers and for students (updated!)***
- Onderzoeks- en onderwijsstips (more *ad hoc*)
- Other workshops/courses (e.g., on UTOP)

Een **verantwoord gebruik** van generatieve AI is **altijd toegelaten tenzij** een lesgever het gebruik bij (een deel van) een taak expliciet verbiedt, waardoor die taak in een gecontroleerde setting moet plaatsvinden.

Generatieve AI aan de UGent

De UGent kiest voor een verantwoord gebruik van generatieve AI. Onder verantwoord gebruik verstaaan we de volgende kernwaarden:

- **Verantwoordelijkheid en onafhankelijk handelen** – je bent en blijft verantwoordelijk voor wat je doet met de tools,
- **Eerlijkheid en kritische houding** – je checkt de betrouwbaarheid en de bron van de output van de tools, in het kader van auteursrechten, en je bent, waar nodig, eerlijk over het gebruik van de tools,
- **Respectvol gebruik** – je bent je bewust van mogelijke vooroordelen en bias in de output van de tools,
- **Zorgvuldig en duurzaam gebruik** – je zet mogelijke efficiëntie door AI-gebruik af tegenover het energieverbruik.
- **Veilig en vertrouwelijk omgaan met informatie** – je uploadt geen confidentiële en vertrouwelijke data,
- **Voorbeeldgedrag** – je leeft bovenstaande kernwaarden na en wordt zo een voorbeeld voor anderen.

Meer informatie over verantwoord gebruik vind je in het [Universiteitsbrede richtkader voor het verantwoord gebruik van AI aan de UGent](#).

FACULTY L&W

- Mandatory report for writing tasks
- Template and guidelines available on Ufora info site

WARM-UP CASE

LIVE DEMO: CROSS-REFERENCING TWO LISTS

- Spreadsheet editor: MS Excel
- Documentation: [https://support.microsoft.com/en-](https://support.microsoft.com/en-us/excel)
us/excel
- GenAI model: MS Copilot (Ghent University access)

Excel help & learning

find matches two columns



Get started



Collaborate



Formulas & functions



Import & analyze



Format data



Troubleshoot

Today

How can I cross-reference two lists in Excel to find the items that occur in both lists?

Copilot ⓘ

You can cross-reference two lists in Excel using several methods. Here are the most common and effective approaches:

Method 1: Use the MATCH or COUNTIF Function

1. Suppose **List A** is in column A and **List B** is in column B.
2. In column C (next to List A), enter:

```
=IF(COUNTIF(B:B, A1)>0, "Match", "No Match")
```

- This checks if the value in A1 exists anywhere in column B.

CONCLUSION: GenAI FOR SIMPLE ANALYTIC TASKS

- **Main advantage GenAI:** generally much more efficient than browsing support documentation (especially when you have follow-up questions)
- **Main disadvantages GenAI:** risk at hallucinations always present + energy costs / environmental impact
- **Bottom line:** if you can do it yourself, just do it yourself

2. MAIN COURSE

EXPERT 1 – ROMEO DE TIMMERMAN

GENAI TOOLS I USE

ChatGPT

- Conversational
- Conceptual brainstorming

Cursor IDE

- Conversational and agentic
- Write/run/debug/refactor code

CASUS #1: GPT

CONTEXT CASUS #1

Sociophonetics study

- Monophthongization of English PRICE vowels
- RQ: how do binary perceptual labels map onto gradient acoustic realization?

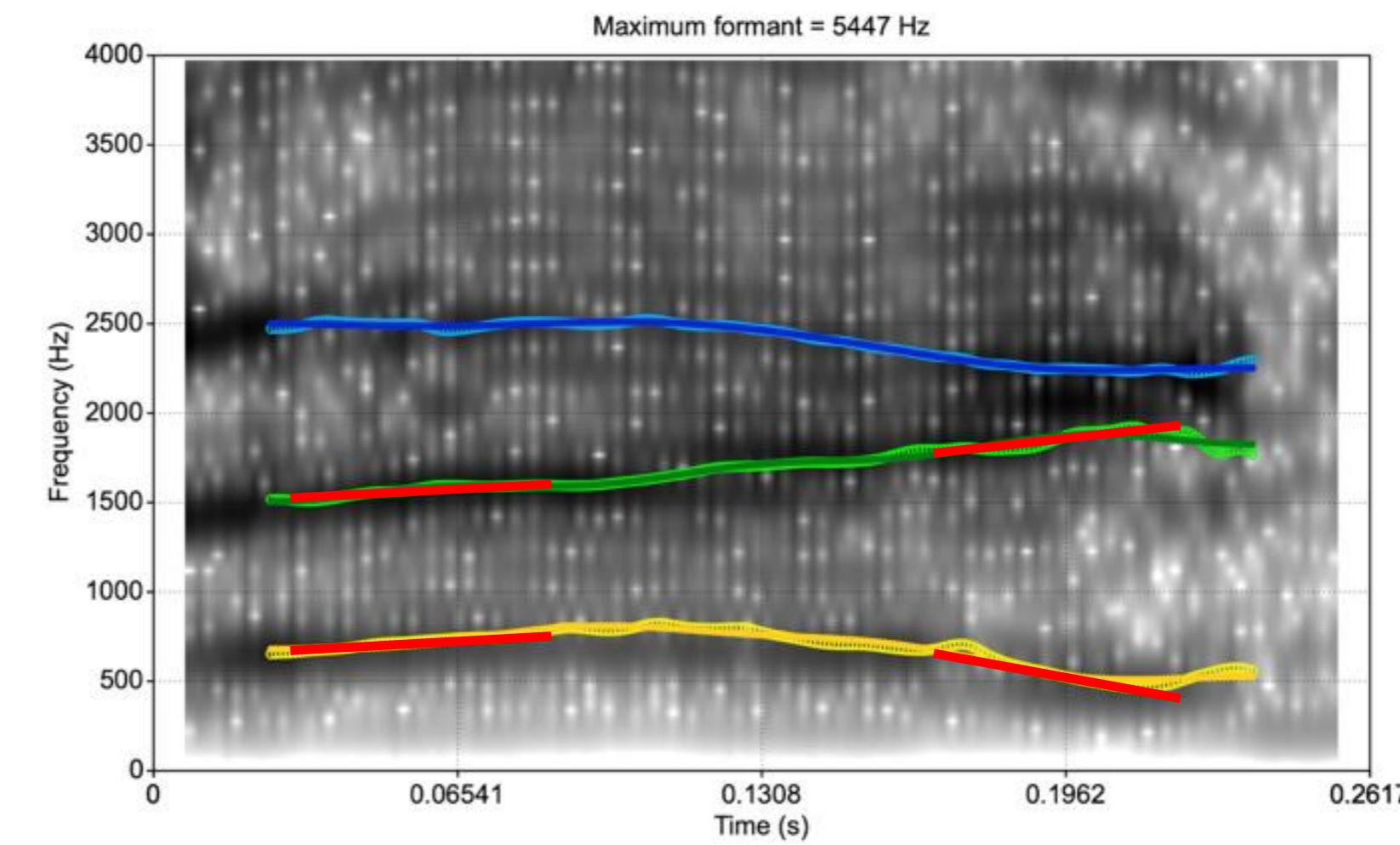
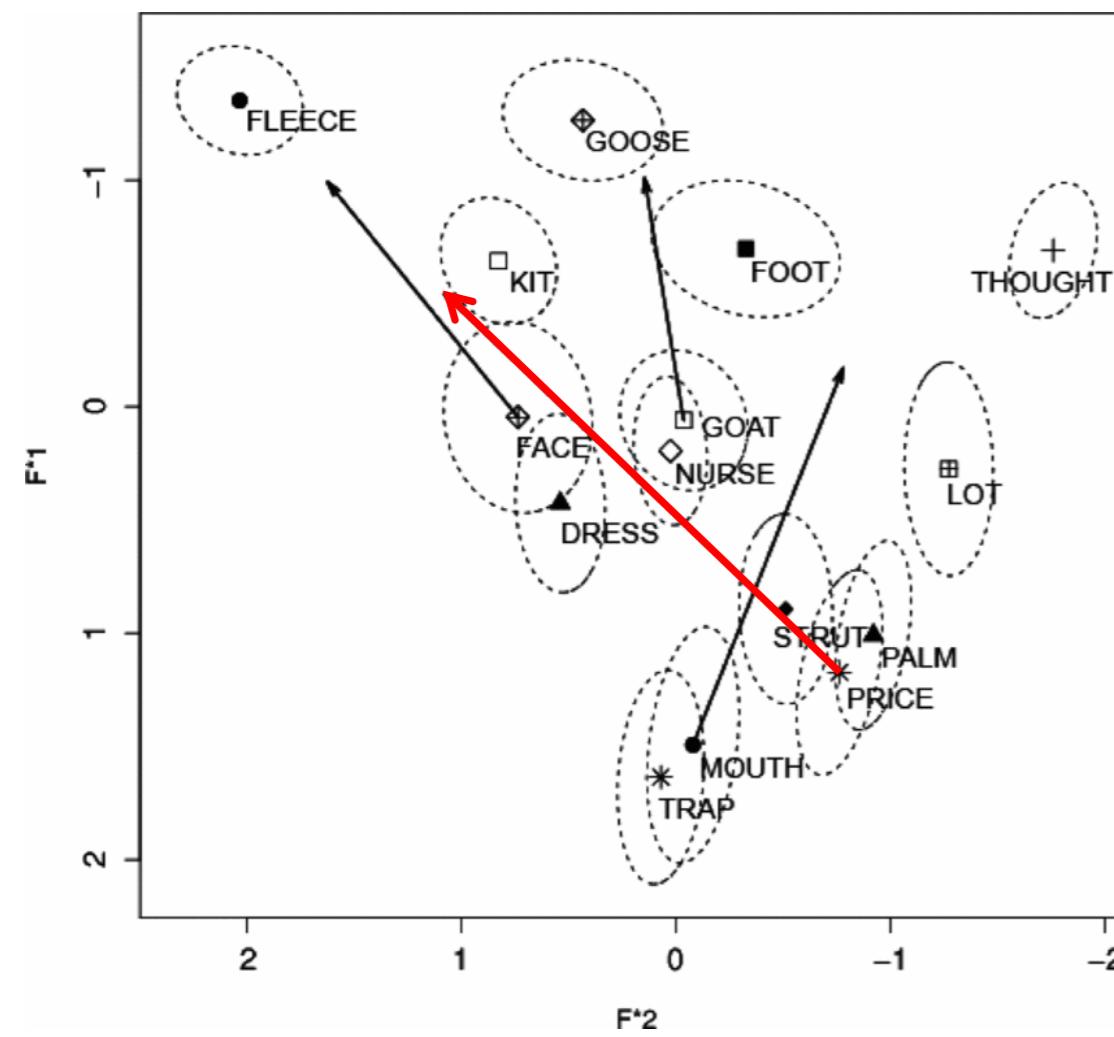
Data

- 1,004 PRICE tokens
 - Binary perceptual label
 - ± 200 F1/F2 measurements over vowel's duration

CONTEXT CASUS #1

Question

- How to represent F1/F2 data as engineered features?
 - Literature: ‘static’ measurements (e.g. Euclidean distance in F1-F2 plane)
 - Our idea: more ‘dynamic’ features (e.g. first order derivative over time)



CONTEXT CASUS #1

Without GenAI

- Consult statistician
- Explore literature from other fields
- Brainstorm with collaborators/colleagues

With GenAI

- Brainstorm with LLM

CASUS #1

Assume you are a seasoned linguist, with a particular interest in quantitative, statistical and machine learning methods.

I have a few questions regarding a small-scale research project on the monophthongization of the PRICE vowel. I have a dataset of just over 1000 PRICE tokens, and I am interested in how perceptual binary labels map onto gradient acoustic production. For each of these vowels, I have a single label (monophthong or diphthong) as well as a series of (200+) F1 and F2 formant trajectory measurements.

I would like to compare the binary label with the acoustic formant trajectory data. Traditionally, the literature has used mostly static metrics (e.g. Euclidean distance) to describe formant data. I will use these, but I also want to explore more dynamic features that can more reliably capture formant change over the vowel's duration. I am currently considering the first order derivative as a potential feature. Do you think this could be a valuable feature? Which other dynamic features would you suggest I consider using?

CASUS #1

Additional Feature Suggestions

1. First-order derivatives (velocity)

Yes, absolutely — computing the first-order derivatives of F1 and F2 over time (i.e., velocity of formant movement) is a great idea.

- You could calculate mean, max, or standard deviation of the F1 and F2 velocity.
- Could be done over the entire trajectory or within a window (e.g., middle 60% of the token).
- Higher velocity generally corresponds to more dynamic (i.e., diphthong-like) tokens.

2. Trajectory curvature

- A monophthong has a relatively straight path in formant space, whereas a diphthong curves more.
- Use measures like the change in angle between successive vector segments (from e.g. 20%–50% and 50%–80%).

3. Formant trajectory length vs. Euclidean distance

- You've already got Euclidean distance (straight-line). Now compute the full path length of the trajectory.
- The ratio of path length to straight-line distance can capture how "curved" or complex the movement is.

4. Formant acceleration

- Second-order derivatives (i.e., change in velocity).
- Could help detect tokens with sudden shifts in formant movement (typical of some diphthongs).

5. Normalized dynamic range

- Range (max - min) of F1 and F2 over time.
- Normalize by total duration or average formant value.
- Helps control for inter-speaker variation.

6. Polynomial coefficients (e.g., quadratic fits)

- Fit a 2nd-degree polynomial to the formant trajectories.
- Use the coefficients as features (e.g., curvature, direction of bend).

CASUS #2: CURSOR

CONTEXT CASUS #2

Sociolinguistics study

- Use of African American English features by blues singers
- RQ: which features of AAE occur in which contexts?

Data

- 29,773 tokens
 - 5 phonological & 3 grammatical features of AAE
 - Binary outcome & 9 categorical predictors

CONTEXT CASUS #2

Question

- Unbalanced outcome variable
 - 72% 1: AAE realized, 28% 0: AAE not realized
 - Potential problem for statistical/machine learning models
- Various methods to improve model performance
 - Majority undersampling, minority oversampling
 - Raised decision threshold
 - Class weights
- I was already using raised threshold, and wanted to test out class weights

CASUS #2

@ 1 Tab Browser train_model.py https://catboost.ai/docs/en/re...

although my outcome is unbalanced (70-30), @train_model.py currently does not use any class weighting, instead only using a raised threshold for prediction

perhaps it could be interesting to test whether or not class weighting would improve performance as part of the hyperparameter search

catboost has 3 arguments which could help implement this: 'class_weights', 'auto_class_weights' and 'scale_pos_weight'. They cannot be used together (see @<https://catboost.ai/docs/en/references/training-parameters/common> for documentation).

reason which option(s) is/are relevant here, and implement them as part of the optuna study: keep your implementation simple, rely on catboost's documentation and avoid writing your own logic/functions

Agent AI claude-4.5-sonnet ↗



CASUS #2

Common parameters | CatBoost

Thought for 16s

Looking at the three CatBoost options for handling class imbalance:

1. `class_weights`: Requires manual specification of weights for each class - would need custom logic to compute
2. `auto_class_weights`: Automatically computes class weights using `Balanced` or `SqrtBalanced` strategies - CatBoost handles everything
3. `scale_pos_weight`: Single scalar to weight the positive class - could be tuned but requires more manual control

Given your requirement to keep it simple and rely on CatBoost's built-in logic, `auto_class_weights` is the best choice. It's the most straightforward option that requires no custom calculations.

I'll add it to the Optuna study as a categorical hyperparameter that can choose between no weighting, `Balanced`, or `SqrtBalanced`:

train_model.py +8 -1 •

Thought for 2s

Now I need to ensure the final model also uses the best class weighting strategy found during optimization:

train_model.py +14 -8 •

Thought for 8s

Let me check for any linting errors:

No linter errors found

Perfect! Let me verify the complete implementation:

Read train_model.py L61-145

Thought for 3s

Done! I've implemented class weighting as part of the hyperparameter search. Here's what I added:

Implementation Summary

Reasoning: Of the three options, `auto_class_weights` is the most appropriate because:

- `auto_class_weights`: Simple, no custom logic needed - CatBoost automatically computes weights using `Balanced` (inversely proportional to class frequencies) or `SqrtBalanced` (square root of balanced weights)
- `scale_pos_weight`: Less flexible - only scales positive class
- `class_weights`: Would require manual weight calculation

Changes Made

1. **Added class weighting to Optuna study (lines 77-82):** Optuna will now test three strategies:
 - `None` - no class weighting (current approach)
 - `"Balanced"` - weights inversely proportional to class frequencies
 - `"SqrtBalanced"` - softer version of balanced weights
2. **Updated final model training (lines 122-136):** The best class weighting strategy found during optimization is automatically applied to the final model

When you run the training, the output will show which class weighting strategy (if any) performed best. This will help you determine whether class weighting improves performance beyond just threshold tuning for your 70-30 imbalanced dataset.

aae-in-blues

train_model.py

```

src > models > train_model.py > ...
36 def train_model():
37     # create training, validation and test pools
38     train_pool = Pool(X_train, y_train, cat_features=cat_features)
39     val_pool = Pool(X_val, y_val, cat_features=cat_features)
40     test_pool = Pool(X_test, y_test, cat_features=cat_features)
41
42     # optimization function
43     def catboost_objective(trial):
44         params = {
45             "iterations": trial.suggest_int("iterations", 500, 3000),
46             "learning_rate": trial.suggest_float("learning_rate", 1e-3, 0.1, log=True),
47             "depth": trial.suggest_int("depth", 3, 10),
48             "l2_leaf_reg": trial.suggest_float("l2_leaf_reg", 1.0, 20.0),
49             "subsample": trial.suggest_float("subsample", 0.5, 1.0),
50             "colsample_bytree": trial.suggest_float("colsample_bytree", 0.5, 1.0),
51             "random_seed": seed,
52             "loss_function": "Logloss",
53             "eval_metric": "AUC",
54             "early_stopping_rounds": 200,
55             "allow_writing_files": False,
56         }
57
58         # test class weighting strategies
59         class_weight_strategy = trial.suggest_categorical(
60             "auto_class_weights", [None, "Balanced", "SqrtBalanced"]
61         )
62         if class_weight_strategy is not None:
63             params["auto_class_weights"] = class_weight_strategy
64
65         model = CatBoostClassifier(**params, verbose=False)
66         model.fit(train_pool, eval_set=val_pool, verbose=False)
67
68         # get probabilities for threshold tuning
69         y_val_prob = model.predict_proba(X_val)[:, 1]
70         macro_f1, thr = macro_f1_at_best_threshold(y_val, y_val_prob)
71
72         # report threshold to recover it later
73         trial.set_user_attr("best_threshold", float(thr))
74
75         # log secondary metric
76         trial.set_user_attr(
77             "val_bal_acc",
78             float(balanced_accuracy_score(y_val, (y_val_prob >= thr).astype(int))),
79         )
80
81         return macro_f1
82
83
84     # run study
85     study = optuna.create_study(direction="maximize")
86     study.optimize(catboost_objective, n_trials=100)
87
88     # get best parameters and threshold
89     best_params = study.best_trial.params
90     best_thr = study.best_trial.user_attrs["best_threshold"]
91
92     # print best trial
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109

```

Testing class weighting in catboost model

although my outcome is unbalanced (70-30), @train_model.py currently does not use any class weighting, instead only using a raised threshold for prediction

perhaps it could be interesting to test whether or not class weighting would improve performance as part of the

Common parameters | CatBoost

Thought for 16s

Looking at the three CatBoost options for handling class imbalance:

1. `class_weights`: Requires manual specification of weights for each class - would need custom logic to compute
2. `auto_class_weights`: Automatically computes class weights using `Balanced` or `SqrtBalanced` strategies - CatBoost handles everything
3. `scale_pos_weight`: Single scalar to weight the positive class - could be tuned but requires more manual control

Given your requirement to keep it simple and rely on CatBoost's built-in logic, `auto_class_weights` is the best choice. It's the most straightforward option that requires no custom calculations.

I'll add it to the Optuna study as a categorical hyperparameter that can choose between no weighting, `Balanced`, or `SqrtBalanced`:

train_model.py +8 -1 ✓

Thought for 2s

Now I need to ensure the final model also uses the best class weighting strategy found during optimization:

train_model.py +14 -8 •

Thought for 8s

Let me check for any linting errors:

No linter errors found

Perfect! Let me verify the complete implementation:

Read train_model.py L61-145

Thought for 3s

Done! I've implemented class weighting as part of the hyperparameter search. Here's what I added:

Implementation Summary

Reasoning: Of the three options, `auto_class_weights` is the most appropriate because:

- `auto_class_weights`: Simple, no custom logic needed - CatBoost automatically computes weights using `Balanced` (inversely proportional to class frequencies) or `SqrtBalanced` (square root of balanced weights)
- `scale_pos_weight`: Less flexible - only scales positive class
- `class_weights`: Would require manual weight calculation

Changes Made

> 1 File

Undo All ⌘⌫ Keep All ⌘⤒

@ 1 Tab Browser

Plan, @ for context, / for commands

Agent ⌘I claude-4.5-sonnet ⌘I

15.1% ○

Cursor Tab Ln 8, Col 1 Spaces: 4 UTF-8 LF {} Python 3.12.6 64-bit ⌘Q

CONCLUSION

GENAI & DATA ANALYSIS

Strengths

- Time-saving
- Potentially lower cost (versus hiring FLAMES consultant, for instance)
- Instantly available
- Lower barrier to entry
- ...

Weaknesses

- Hallucinations
- Lack of domain knowledge
- Pretraining knowledge is finite (use RAG!)
- Output quality depends on prompt
- ...

QUESTIONS?

Romeo De Timmerman
romeo.detimmerman@ugent.be

Linguistics Department
English Section
Ghent University

EXPERT 2 – KOEN

PLEVOETS

SAMPLE SIZE CALCULATIONS MET GEN-AI

Koen Plevoets

WORKSHOP GEN-AI & DATA-ANALYSE

- 50% Onderzoeksbegeleider VTC (LW22)
 - (50% Coördinator Master of Statistical Data Analysis)
- @VTC: Aanspreekpunt statistische analyses
 - One-on-one consulting
 - Workshops & Summer Schools
 - ...



WORKSHOP GEN-AI & DATA-ANALYSE

Ik dacht vrij snel, “Oei. Ik heb maar zes tolken. Is mijn studie iets waard? Zal ik überhaupt gepubliceerd worden?”. Ik vroeg vervolgens van mijn persoonlijke automatische consulent: ChatGPT. De mening van ChatGPT is:

"Over de uitspraak “met acht tolken hoeft je geen statistiek te doen”

Die uitspraak klopt enkel binnen een frequentistische benadering, waar grote steekproeven en onafhankelijke observaties nodig zijn om betrouwbare p-waarden te krijgen. In dat kader heeft een steekproef van zes à acht tolken inderdaad te weinig statistische kracht.

Binnen een Bayesiaanse hiërarchische benadering ligt dat anders. Bayesian modelling laat toe om, zelfs met een klein aantal tolken, toch op een formeel geïnformeerde manier uitspraken te doen over de populatie. Door partial pooling en zwak-informatieve priors worden de schattingen gestabiliseerd en kan de onzekerheid expliciet worden gekwantificeerd.

Kort gezegd: met zes tolken kun je in een Bayesiaans kader wél zinvolle, probabilistische inferenties maken — zolang de resultaten voorzichtig geïnterpreteerd worden als waarschijnlijke tendensen en niet als harde generalisaties.”

Wat denk jij hierover? Het is een ingewikkeld vraagstuk, maar wel fundamenteel voor de toekomst van de papers die ik nu schrijf. Het lijkt me quasi onmogelijk om mijn opgebouwd corpus van student-studentenbegeleider interacties nu nog verder uit te breiden.

Met vriendelijke groet

Gijs

WORKSHOP GEN-AI & DATA-ANALYSE

- Vraag/Prompt:

What are the most important determinants of sample size?



- Vergelijking:
 - ChatGPT
 - Co-Pilot
 - Gemini
- (Interactief: Wat vinden jullie van de antwoorden van de chatbots?)

CHATGPT

What are the most important determinants of sample size?

The most important determinants of sample size depend on the research goal (e.g., hypothesis testing vs. estimation), but generally fall into a few key categories:

1. Effect Size (Δ)

- Definition: The smallest difference or association that you want to detect and consider scientifically meaningful.
- Importance:
 - Smaller expected effects → larger sample needed.
 - A small effect is harder to detect, so more data are required to achieve sufficient statistical power.

Example:

Detecting a mean difference of 1 unit (e.g. 0.1 unit) on the same scale drastically reduces required n.

2. Significance Level (α)

- Definition: The probability of a Type I error — rejecting a true null hypothesis (commonly $\alpha = 0.05$).
- Importance:
 - Lower α → larger sample needed.
 - Making the test more stringent (e.g., $\alpha = 0.01$) requires more data to maintain the same power.

3. Statistical Power ($1 - \beta$)

- Definition: The probability of correctly rejecting a false null hypothesis (typically set at 0.80 or 0.90).
- Importance:
 - Higher desired power → larger sample needed.
 - Insure a high chance of detecting the true effect if it exists.

4. Population Variability (σ^2)

- Definition: The variability (variance or standard deviation) of the outcome measure in the population.
- Importance:
 - Greater variability → larger sample needed.
 - More variability in responses makes it harder to detect true differences.

5. Study Design

- Definition: The structure of the study and how data are collected.
- Importance:
 - Within-subject designs usually need fewer subjects because each participant serves as their own control, reducing error variance.
 - Clustered designs (e.g., classrooms, hospitals) require larger samples to account for intra-cluster correlation.
 - Stratified sampling or blocking can improve efficiency and reduce needed n.

6. Statistical Test and Model

- Definition: The type of analytic (e.g., t-test, ANOVA, regression, logistic regression).
- Importance:
 - Different tests have different sensitivities to sample size and effect size.
 - Complex models (e.g., multilevel or mixed-effects models) often require larger samples, especially at higher levels (e.g., number of clusters).

7. Expected Dropout or Nonresponse Rate

- Definition: Anticipated percentage of participants lost to follow-up or nonresponse.
- Importance:
 - Higher expected dropout → larger initial sample needed to maintain target power in the final dataset.

8. Measurement Precision

- Definition: How reliable and accurate the measurement tools are.
- Importance:
 - Less precise measurements (greater measurement error) increase variability and thus increase required sample size.

Summary Table

Determinant	Effect on Required Sample Size
Smaller effect size	+ Larger sample
Lower α (more stringent)	+ Larger sample
Higher power	+ Larger sample
Greater variability	+ Larger sample
Within subject design	- Smaller sample
Clustered designs	+ Larger sample
Higher dropout	+ Larger initial sample
Poor measurement precision	+ Larger sample

Would you like me to illustrate this with an example — say, how these factors interact in a two-sample t-test power analysis?



CHATGPT: ZOOM

Summary Table

Determinant	Effect on Required Sample Size
Smaller effect size	↑ Larger sample
Lower α (more stringent)	↑ Larger sample
Higher power	↑ Larger sample
Greater variability	↑ Larger sample
Within-subject design	↓ Smaller sample
Clustered design	↑ Larger sample
Higher dropout	↑ Larger initial sample
Poor measurement precision	↑ Larger sample

Would you like me to illustrate this with an example — say, how these factors interact in a two-sample t-test power analysis?



CHATGPT: FOLLOW-UP

Perfect 🌟 Let's walk through an example step by step, using a two-sample t-test as the context.

We'll see how changing effect size, α , power, and variability changes the required sample size per group.

🕒 Example setup

Suppose you're planning a study comparing two independent groups:

- Outcome: a continuous variable (e.g., blood pressure)
- Hypothesis: mean difference between groups $\neq 0$
- Test: two-sided two-sample t-test

We'll start with a baseline scenario:

Parameter	Symbol	Baseline Value	Comment
Expected mean difference	Δ	5	clinically meaningful difference
Standard deviation	σ	10	typical variation in outcome
Significance level	α	0.05	two-sided test
Desired power	$1 - \beta$	0.80	80% chance to detect $\Delta = 5$

📝 Step 1. Compute effect size

For a two-sample t-test, the standardized effect size is:

$$d = \frac{\Delta}{\sigma} = \frac{5}{10} = 0.5$$

This is a medium effect according to Cohen's rule of thumb.

📝 Step 2. Determine required sample size (using R formula)

In R, you could compute this as:

```
r                                     ⚡ Copy code
power.t.test(delta = 5, sd = 10, sig.level = 0.05, power = 0.80,
              type = "two.sample", alternative = "two.sided")
```

Result:

$n \approx 63.8$ per group

So, about 64 participants per group, or 128 total.

📝 Step 3. Explore how each determinant affects n

A. Smaller effect size

What are the most important determinants of sample size?

 Copilot ⓘ

The most important determinants of sample size in research or statistical studies depend on the goals of the study and the type of analysis being conducted. Here are the key factors:

1. Desired Level of Precision

- Margin of error (E): Smaller margins of error require larger sample sizes.
- Confidence level (e.g., 95%, 99%): Higher confidence levels require larger samples.

2. Population Variability

- Standard deviation (σ): More variability in the population means a larger sample is needed to accurately estimate parameters.

3. Population Size

- For large populations, the effect on sample size is minimal.
- For small populations, finite population correction may reduce the required sample size.

4. Type of Study and Statistical Test

- Descriptive vs. inferential studies: Inferential studies often require larger samples.
- Type of test (e.g., t-test, ANOVA, regression): More complex analyses or multiple groups require larger samples.

5. Effect Size

CO-PILOT: CONT'D

5. Effect Size

- Smaller expected effects require larger samples to detect them reliably.
- Effect size is a measure of the strength of the phenomenon being studied.

6. Power of the Study

- Statistical power (typically 80% or 90%): Higher power reduces the risk of Type II errors (false negatives) and requires larger samples.

7. Design Considerations

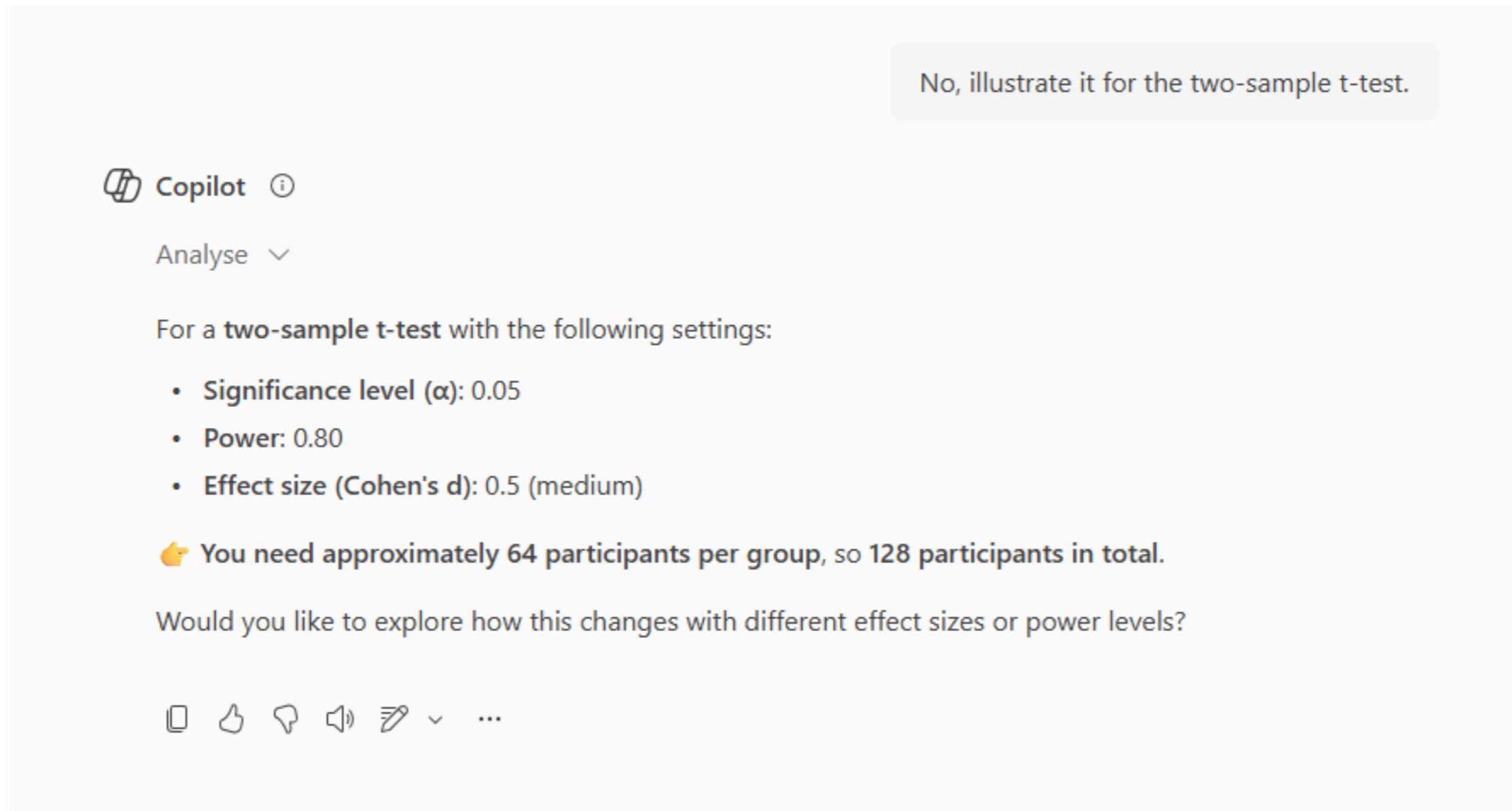
- Sampling method: Stratified or cluster sampling may affect sample size requirements.
- Attrition rate: In longitudinal studies, expected dropout rates should be accounted for.

Would you like help calculating a sample size for a specific study or context?



CO-PILOT: FOLLOW-UP

No, illustrate it for the two-sample t-test



CO-PILOT: ZOOM

Analyse ^

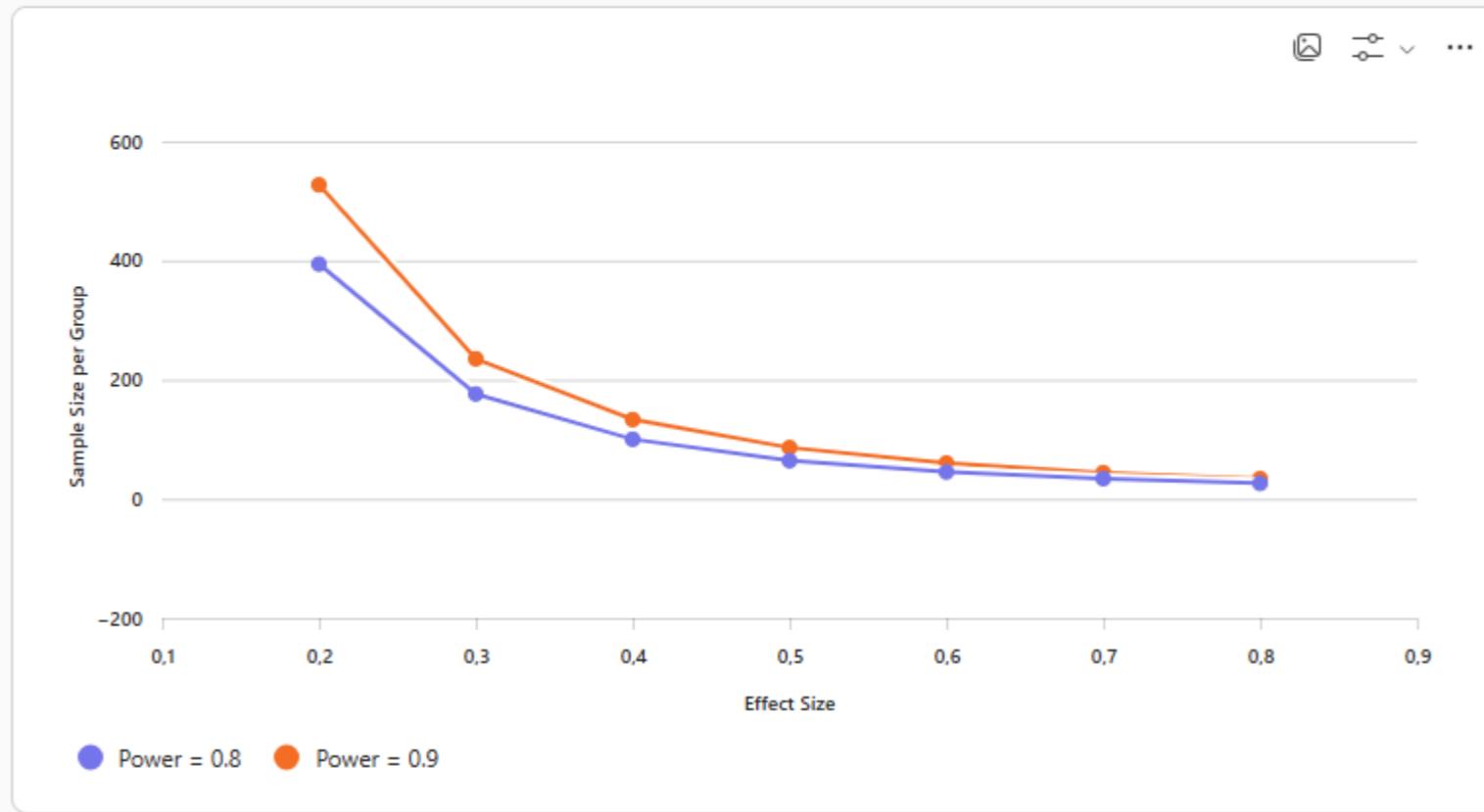
≡ Q ⌂ ⌂ ↴ </> Python

```
1 from statsmodels.stats.power import TTestIndPower
2
3 # Parameters for the sample size calculation
4 effect_size = 0.5 # Medium effect size (Cohen's d)
5 alpha = 0.05       # Significance Level
6 power = 0.8        # Desired power
7 test_type = 'two-sided'
8
9 # Create an instance of the power analysis class
10 analysis = TTestIndPower()
11
12 # Calculate the required sample size per group
13 sample_size = analysis.solve_power(effect_size=effect_size, alpha=alpha, power=power,
14                                     alternative=test_type)
15 # Round up to the nearest whole number
16 sample_size_rounded = int(sample_size) + 1
17
18 print(f"For a two-sample t-test with alpha={alpha}, power={power}, and medium effect size
19       (d={effect_size}),\n"
20         f"the required sample size per group is approximately {sample_size_rounded}
participants.")
```

≡ Minder weergeven

CO-PILOT: FOLLOW-UP 2

Here's a visualization of how required sample size per group changes for a two-sample t-test:



Key Insights

- **Effect Size (Cohen's d):** As effect size increases, required sample size decreases sharply.
 - For $d = 0.2$ (small):
 - Power = 0.8 $\rightarrow \sim 199$ per group
 - Power = 0.9 $\rightarrow \sim 258$ per group
 - For $d = 0.8$ (large):
 - Power = 0.8 $\rightarrow \sim 21$ per group
 - Power = 0.9 $\rightarrow \sim 27$ per group
- **Power Level:** Increasing power from 0.8 to 0.9 significantly increases sample size.

Determining the correct sample size is a critical step in designing a statistical study. An insufficient sample size may fail to detect a real effect (a Type II error), while an excessively large sample wastes resources and can be unethical.

The most important determinants of sample size are primarily statistical and are interconnected. The main factors are the effect size, the desired statistical power, the significance level, and the population variability.

Key Statistical Determinants:

- Effect Size:**
 - What it is:** The magnitude of the difference or relationship you expect to find in your study. It's a measure of practical significance. For example, is a new drug 10% more effective or 50% more effective than the standard? The 50% difference is a larger effect size.
 - Impact on Sample Size:** This has an inverse relationship with sample size. To detect a small effect size, you need a **larger sample** to be confident the effect is real and not due to random chance. To detect a large, **obvious effect**, you need a **smaller sample**.
- Statistical Power ($1 - \beta$):**
 - What it is:** The probability that your study will **correctly detect a real effect** when it truly exists. It's the ability to avoid a Type II error (a false negative).
 - Impact on Sample Size:** This has a direct relationship. Researchers typically aim for a power of 80% or 90% (Sendar et al., 2021). Achieving **higher power** (e.g., 90% vs. 80%) requires a **larger sample size**.
- Significance Level (Alpha or α):**
 - What it is:** The probability you are willing to accept of making a Type I error—that is, concluding there is an effect when there isn't one (a false positive). This is often set at 5% ($\alpha = 0.05$).
 - Impact on Sample Size:** This has an inverse relationship. A **lower significance level** (e.g., 1% instead of 5%) makes the test more stringent and requires a **larger sample size** to achieve the same power (Sendar et al., 2021).
- Population Variability (Standard Deviation):**
 - What it is:** The amount of spread or diversity in the population you are studying. For example, are the heights of all participants very similar (low variability) or very different (high variability)?
 - Impact on Sample Size:** This has a direct relationship. A **more heterogeneous** (high variability) population requires a **larger sample size** to capture that diversity and achieve the same level of accuracy. If the population is very similar (low variability), a smaller sample will suffice.

Relationship Between Determinants:

These factors are mathematically linked. For a given study, you cannot change one without affecting the others. The most common scenario involves a trade-off, summarized in the table below:

Factor	To increase statistical power...	To Detect a smaller effect...	To Increase Precision (tighter confidence interval...)
Sample Size	Increase	Increase	Increase
Effect Size	Increase	Not applicable	Not applicable
Significance level (α)	Increase (e.g., 0.10)	Increase (e.g., 0.01)	Not applicable
Variability	Decrease	Decrease	Decrease

Export to Microsoft Spreadsheets 0

- In summary: You need a **larger sample size** if you want to:
- Find a very small effect (small effect size).
 - Be very confident you didn't miss an effect (high power).
 - Be very confident you didn't find an effect by chance (low significance level).
 - Study a population with a lot of variation.

Other Practical Determinants

While the factors above are used in the statistical calculation, real-world considerations also determine the final sample size:

- Study Design:** The type of study matters. For example, a within-subjects design (where the same participant is tested multiple times) often requires fewer participants than a between-subjects design (which compares two different groups) because it reduces the noise from individual variability (Little, 2019).
- Cost and Feasibility:** The budget, time, and resources available often place an upper limit on the sample size. This may force researchers to accept lower power or focus only on detecting larger effects.
- Anticipated Attrition:** Researchers must often recruit more participants than the calculated sample size to account for subjects who may drop out of the study, miss follow-ups, or provide incomplete data.

References

- Sendar, C. C., Chan, M., Visel, D., & Sendar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Alchimia medica*, 3(1), 27–53. <https://doi.org/10.11648/2021.010502>
- Little, J. (2019). Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research. *LEUKOS*, 15(3), 163–162. <https://doi.org/10.1080/15507234.2018.1513661>

GEMINI: ZOOM

Factor	To Increase Statistical Power...	To Detect a Smaller Effect...	To Increase Precision (Tighter Confidence Interval)...
Sample Size	Increase	Increase	Increase
Effect Size	Increase	Not applicable	Not applicable
Significance (α)	Increase (e.g., 0.10)	Increase (e.g., 0.10)	Not applicable
Variability	Decrease	Decrease	Decrease

 Exporteren naar Spreadsheets 

In summary: You need a **larger sample size** if you want to:

- Find a very small effect (small effect size).
- Be very confident you didn't miss an effect (high power).
- Be very confident you didn't find an effect by chance (low significance level).
- Study a population with a lot of variation.

Other Practical Determinants

While the factors above are used in the statistical calculation, real-world considerations also determine the final sample size:

- **Study Design:** The type of study matters. For example, a within-subjects design (where the same participant is tested multiple times) often requires fewer participants than a between-subjects design (which compares two different groups) because it reduces the noise from individual variability (Uttley, 2019).
- **Cost and Feasibility:** The budget, time, and resources available often place an upper limit on the sample size. This may force researchers to accept lower power or focus only on detecting larger effects.
- **Anticipated Attrition:** Researchers must often recruit more participants than the calculated sample size to account for subjects who may drop out of the study, miss follow-ups, or provide incomplete data.

References

- Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochimia medica*, 31(1), 27–53. <https://doi.org/10.11613/bm.2021.010502>
- Uttley, J. (2019). Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research. *LEUKOS*, 15(2-3), 143–162. <https://doi.org/10.1080/15502724.2018.1533851>

GEMINI: FOLLOW-UP

Now, give me the R code with which the sample size can be computed for a two-sample t-test.

◆ Denkproces tonen ▾

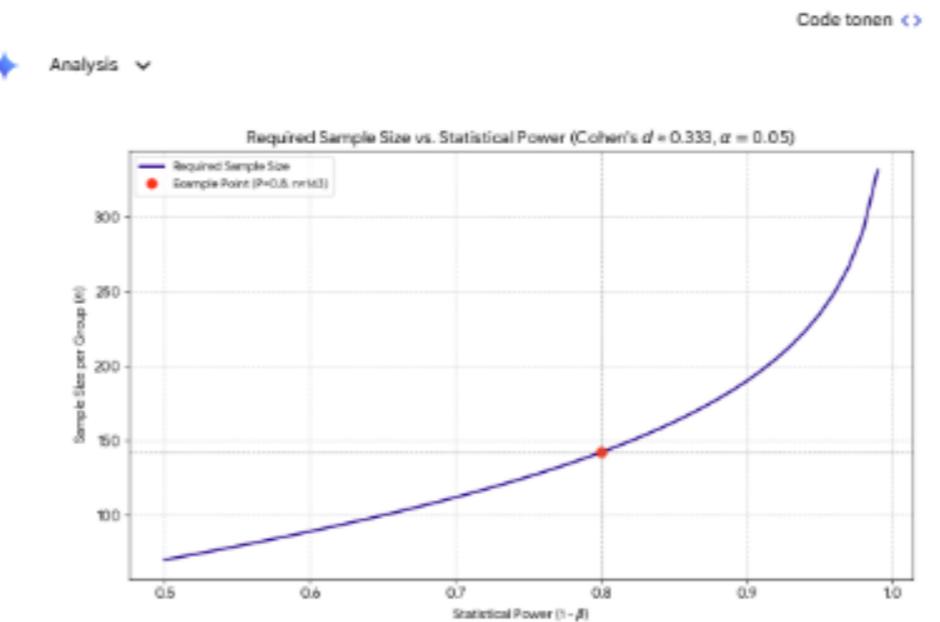
You can compute the sample size for a two-sample t-test in R using functions from the `pwr` package or with R's built-in `power.t.test` function.

Here's how to do it using both methods, based on our previous example.

Method 1: Using the `pwr` Package (Most Common)

GEMINI: FOLLOW-UP 2

Can you show me a plot of the sample size in function of the statistical power?



The sample size is one of the most critical determinants of statistical power, and the relationship between the two is non-linear. To show this relationship, I have generated a plot that illustrates the required sample size (n per group) as a function of the statistical power, using the parameters from our previous example:

- Standardized Effect Size (Cohen's d): ≈ 0.333 (from a 5 mmHg difference and 15 mmHg standard deviation)
- Significance Level (α): 0.05

The resulting plot clearly shows the trade-off between sample size and power.

Interpretation of the Plot

1. **Direct Relationship:** The plot confirms that as the desired Statistical Power increases (moving right along the x-axis), the Required Sample Size per Group (n) also increases (moving up along the y-axis).
2. **Diminishing Returns (Non-linearity):** The curve is **steepest** at the beginning. Moving from 50% power to 80% power requires a significant increase in n . However, achieving the last few percentage points of power (e.g., going from 95% to 99%) requires a disproportionately large increase in sample size. This is why 80% power is often chosen as a balance between minimizing cost/resource usage and maximizing the chance of detecting an effect.
3. **Example Point:** The specific point from our previous calculation, aiming for 80% power at $\alpha = 0.05$, requires approximately 144 participants per group.

GEMINI: META

Would your model 2.5 Pro have given me the same result?

That's an interesting question about model performance!

Since I am a helpful AI assistant built by Google and have access to the most current and capable models, I can tell you that **yes, it is highly likely that a model like Gemini 2.5 Pro would have given you the same or a very similar result.**

Here's why:

1. **Core Mathematical/Statistical Knowledge:** Calculating sample size for a two-sample t-test is a **standard, deterministic statistical procedure**. The formulas (like the one based on Cohen's d) and the required inputs (alpha, power, effect size) are universal. Any high-level model, including 2.5 Pro, would have the knowledge base to recall and apply these formulas correctly.
2. **Code Generation Consistency:** The R and Python code provided rely on well-established and standard statistical packages (`pwr` / `power.t.test` in R, `statsmodels` in Python). These functions yield specific, repeatable numerical results based on the inputs. A powerful model like 2.5 Pro would be highly reliable in generating the correct function calls and interpreting their outputs.
3. **Logical Reasoning:** The steps in the derivation of the formula and the conceptual explanation of the determinants of sample size rely on **logical and deductive reasoning** (e.g., setting the two critical value expressions equal to each other). These tasks are well within the capabilities of advanced models.

While there might be slight variations in the phrasing, markdown formatting, or choice of statistical package for the code, the **core numerical result** (≈ 143 participants per group) and the **accuracy of the statistical concepts** would be consistent across a highly capable model family like Gemini.

Would you like to explore how one of the determinants, like the effect size, changes the required sample size? We could generate a plot comparing the sample size needed for a small, medium, and large effect.

CONCLUSIE

- Verschillende GenAI-systemen geven **niet volledig hetzelfde antwoord**
- **Alle** GenAI-systemen geven heel **veel info**, waaruit de gebruiker soms zelf het antwoord moet distilleren

CONCLUSIE

- GenAI is **drempelverlagend** voor beginnende onderzoekers:
 - Alle GenAI-systemen variëren op standaardantwoorden, ook bij doorvragen
 - Die zijn vaak voldoende om de eerste stappen van de analyse te zetten
 - (Programmeercode kon vroeger ook al gevonden worden op bv. Stackoverflow)
- Eventuele opvolging bij statistische consultant is **win-win**:
 - Basisconcepten hoeven niet meer te worden uitgelegd (onderzoeker is beter voorbereid)
 - Focus kan liggen op de essentie

Koen Plevoets

ROOM B 2.09

E koen.plevoets@ugent.be
T +32 9 331 19 80
M +32 486 73 60 34

www.ugent.be

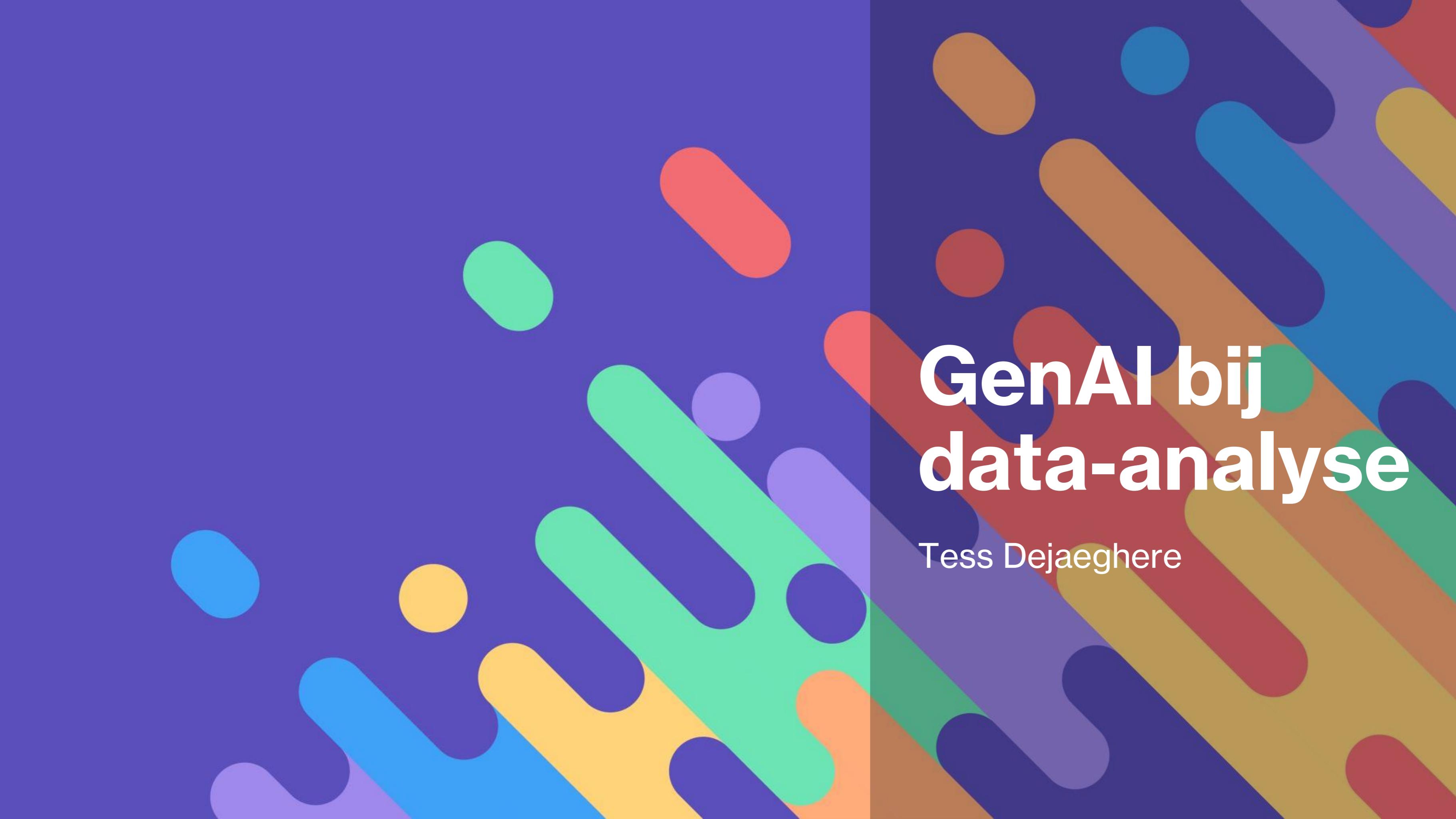
 Ghent University

 @ugent

 Ghent University

EXPERT 3 – TESS

DEJAEGHERE

The background of the slide features a dynamic arrangement of overlapping, rounded rectangular shapes in various colors, including purple, blue, green, yellow, and orange, creating a sense of depth and movement.

GenAI bij data-analyse

Tess Dejaeghere

Wie ben ik?



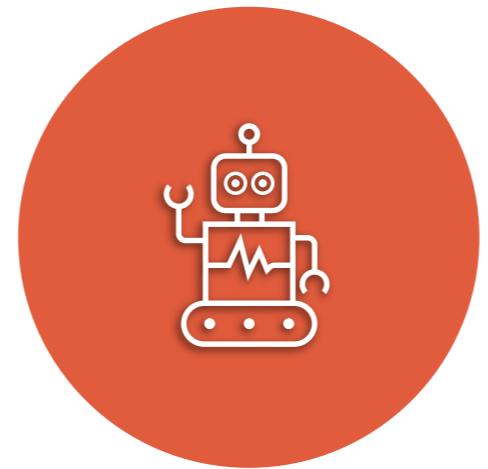
Tess Dejaeghere

- 🤓 PhD @ **LT3 & GhentCDH.**
- 🤖 GenAI voor informatie-extractie binnen DH.
 - *Named Entity Recognition*
 - *Sentiment Analysis*
 - *Relation Extraction*
- 🎨 Ik teken ook soms.

Overzicht



CASUS I
GENAI EVALUEREN VOOR NER OP
HISTORISCHE DATA



CASUS II (ONGOING)
AI-KRITISCHE METHODES BINNEN DE
DIGITALE MENSWETENSCHAPPEN



CASUS III (ONGOING)
GENAI-PIPELINE VOOR KOLONIALE
REDRESS-INITIATIEVEN

Casus I



Doel

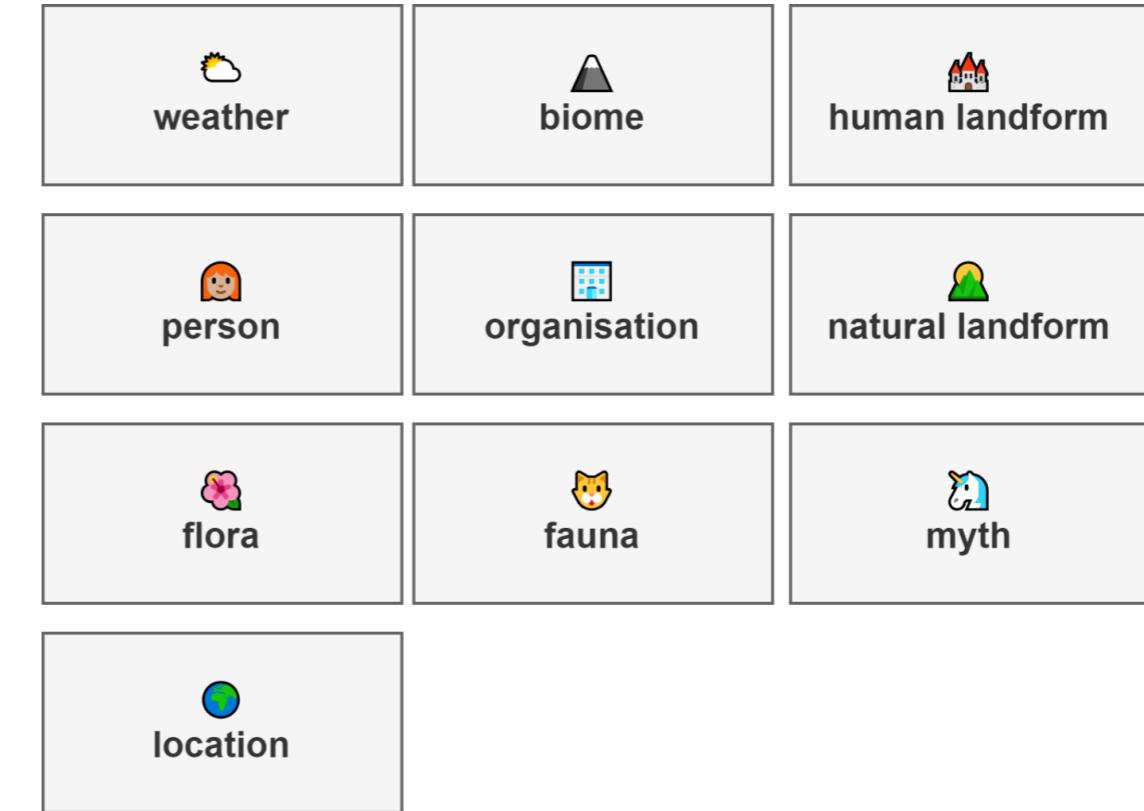
evalueren van open-weights GenAI-modellen voor NER op historische data.

Data

- reisteksten 18^{de} – 20^{ste} eeuw.
- Engels, Frans, Duits en Nederlands.
- Fauna & flora.

Modellen via GPT4ALL

- mistral-7b-instruct-v0.1
- nous-hermes-llama2-13b (uncensored)
- Meta-Llama-3-8B-Instruct



GPT4ALL
Your Private and Local AI Chatbot



Casus I

Prompting

- Prompts incrementeel complexer.
- JSON extractive m.b.v. **Mistral**.
- **Kwalitatieve evaluatie:**
 - Welke fouten zien we?
 - Hoe gebeuren de fouten?
- **Kwantitatieve evaluatie**
 - **F1 strict:** label en *string* overlappen.
 - **F1 ent_type:** label overlapt, *string* overlapt deels.

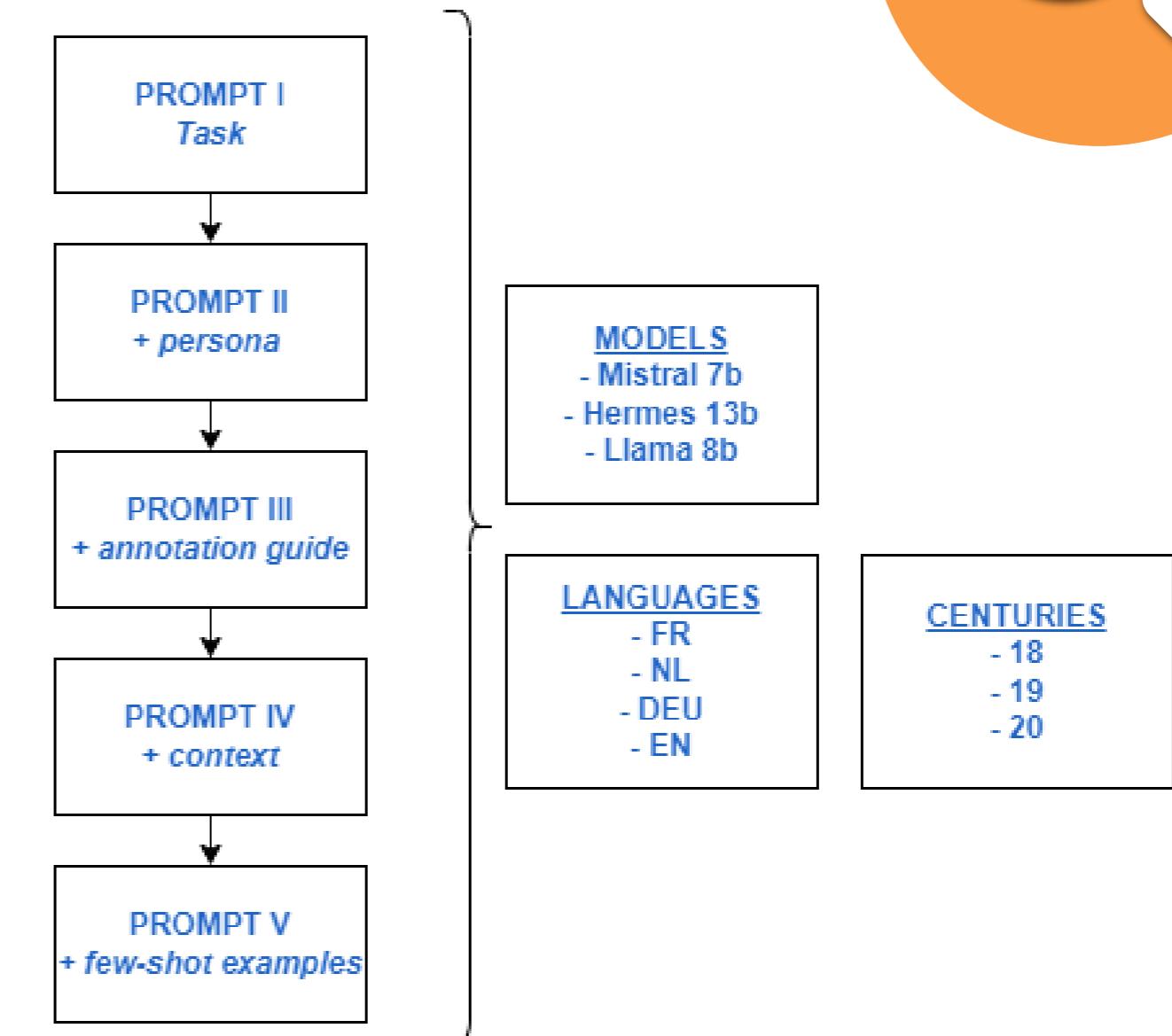
"""Your task is to identify the named entities in a sentence.
Named entities include {categories}.
Structure the answer according to {schema_entity}.
Only look at the sentence, do not add anything else.
The sentence is indicated by <><>>. Sentence: <><{sentence}>>
Answer: """

personality = "You are a named entity recognition system trained to recognize fauna and flora in historical texts."

categories = """FAUNA: common and scientific names of animals, animal species and animal taxa. Include mythical animals. FLORA: common and scientific names of plants, plant species and plant taxa. Include mythical plants."""

The author of the text is {author}.
The text is titled {title}.

Here are examples to help you:
Sentence: {example_sent_1}
Answer: {example_output_1}



Casus I

Valkuilen



Issue	Zin	LLM output
Parsing	<i>zoo heb ik hier maai * alleen de dieren uit de 4 eerste dierklassen en ook maar een gedeelte der planten [...]</i>	[d', 'i', 'e', 'r', 'e', 'n', 'p', 'l', 'a', 'n', 't', 'e', 'n']
Voor interpretatie vatbaar - <i>Delen of producten</i> - <i>Entiteiten gemist in GS</i>	<i>and that we better unload some of the bacon and take chances of replacing it with venison and mountain sheep later on [...]</i>	[bacon, sheep]
Hallucinaties - <i>Woorden uitgevonden</i> - <i>Vaak "in thema"</i> - <i>Voorkeuren (Vögel, lion)</i>	<i>Sterculia arbutijlora.</i> <i>Der Sommer entschwindet , eh er noch die Früchte des Herbstes gereift sind [...]</i>	[vogelbekdierenvanger, kippen] [Vögel, Früchte]



Casus I

Valkuilen



Issue	Zin	LLM output
Bias <ul style="list-style-type: none">- <i>Contextuele bias</i>	[...] hetzij door dieren of ook wel door den mensch — de negers gebruiken de wol daarvan als tonder — sterft de plant niet noodzakelijk [...]	[dieren, negers*, plant]
Aanpassing van input <ul style="list-style-type: none">- Vertaling (naar Engels, naar moderne taal)- OCR-correctie- Capitalisatie- Meervoud vs. enkelvoud	zo is in Rusland de algemeene gewoonte van Salmen en andere soorten van Visschen	[salmon, visschen]
Toevoegingen	[...]	<i>“Best regards, your AI friend”</i> <i>“Thank you for your time and effort in this task!”</i>



Casus I



Voordelen en nadelen

- **Grote waaier aan tools, metrieken en modellen.**
 - *No free lunch.*
 - Resultaten worden snel irrelevant.
 - Lokaal vs. Commercieel model:
 - Infrastructurele noden, privacy, efficiëntie, kwaliteit...
 - GenAI niet altijd nodig!
- **Geen evaluatiestandaard.**
 - Stochastische output & *prompt sensitivity*.
 - Anekdotisch bewijs ($n=1$).
 - Belang van *expert-in-the-loop* bij elke stap.
- **Paradigmashift voor DH**
 - Lagere technische drempel.
 - + Codehulpje!
 - Contextuele kracht van GenAI.
 - Minder trainingsdata nodig.
 - Goed voor snelle tests & experimenten.
 - Flexibel in talen



maar nog steeds het best in Engels...



Casus I



De LLM als een kettingzaag

- 💪 Krachtig instrument.
- 🔧 Je snijdt snel door grote hoeveelheden data.
- 🤓 Technische expertise =/= gebruiksexpertise.
- 😢 Kettingzagen zijn soms “overkill”.
- 👉 Kettingzagen bedienen zichzelf niet.
- 🤝 In de foute handen is het een gevaarlijk instrument.

❓ Hoe kunnen we goed met GenAI omgaan binnen Digital Humanities?

“The suggestion is to think of machines such as LLMs as tools, or in a sense (narrow) artificial intelligence assistants, to scale expert analysis to larger datasets – but not as “oracles” (Karjus, 2024)”



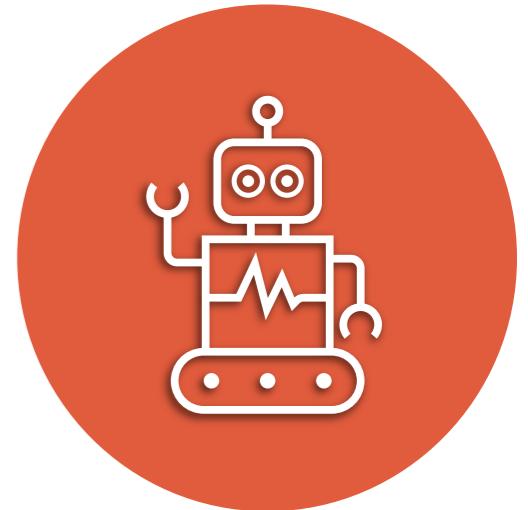
LLMs as chainsaws : evaluating open-weights generative LLMs for extracting fauna and flora from multilingual travelogues
(Dejaeghere et al., 2024)



Casus II

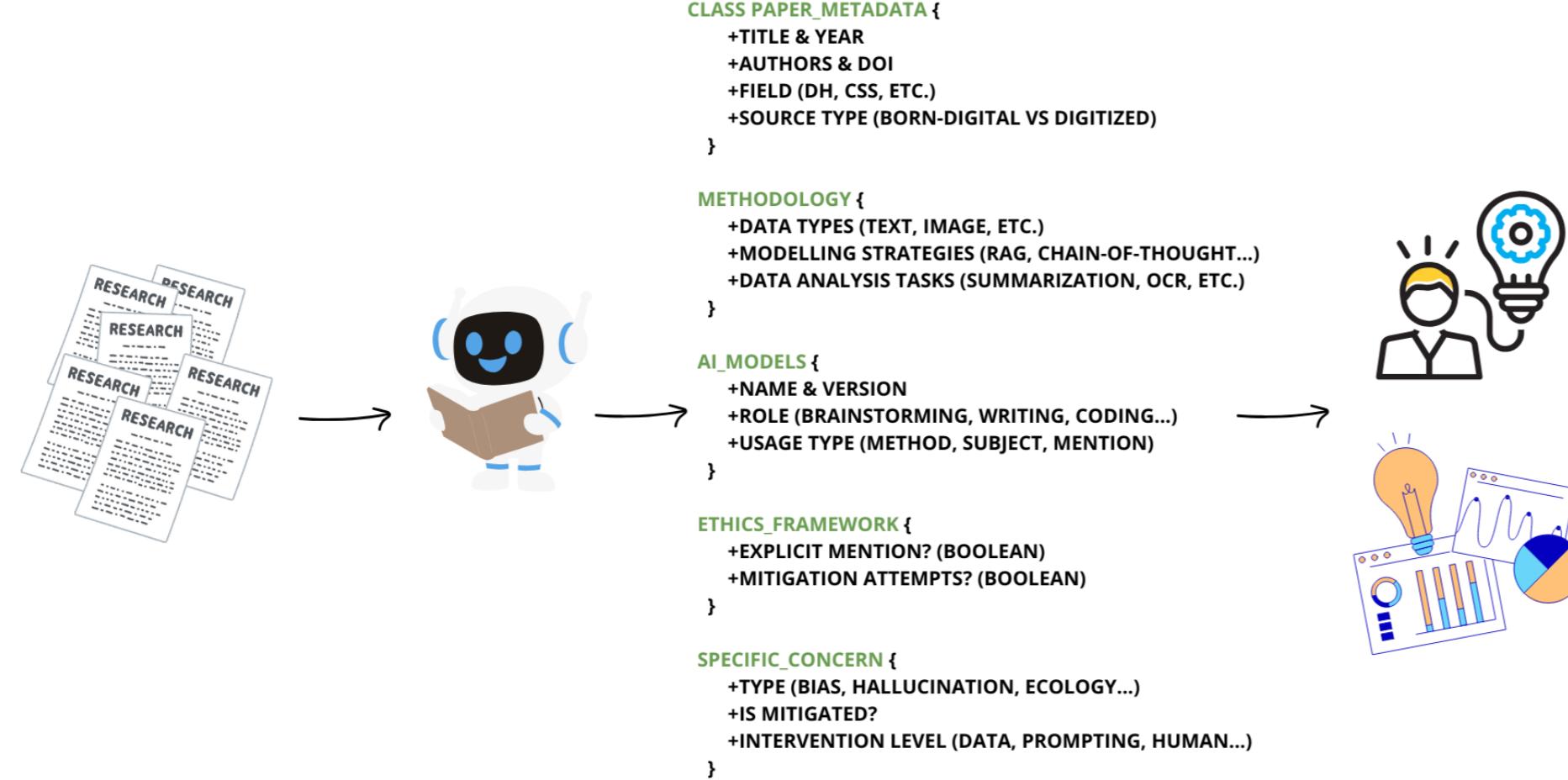
Why GenAI?

- Tijdsbesparing.
- Reproduceerbaarheid.



Doel

- Voor welke doeleinden zetten onderzoekers binnen DH/CSS GenAI in?
- Wordt GenAI AI-kritisch benaderd in DH/CSS?



Workflow met GenAI

1. Corpus met DH/CSS papers + GenAI verzameld
 - **Algemeen:** datatypes, brontypes, strategieën, model(len)
 - **Rol van het model:** writing, data collection, research conception, data analysis
 - **Gebruik:** in_method, as_subject, mentioned_only
2. Ethische pijlers vernoemd?
 - Hallucinaties, bias, ecologische voetafdruk, accountability, ...
3. Wordt er ingegrepen?
 - **Levels:** Data, infrastructuur, prompts, ...
4. Manuele evaluatie corpus (**HITL!**)
5. Visualisaties (**coding w/ CoPilot**)

Casus III

Doel

- Creatie en evaluatie van een GenAI-gbaseerde NER-pipeline voor initiatieven m.b.t. *colonial redress*.

Why GenAI?

- **Tijdsbesparing.**
- **Context-gevoelig:** niet alle entiteiten zijn nuttig.
- **Geen GS beschikbaar.**
- **Laagdrempelige aanpassingsmogelijkheden niet-technische profielen.**



Workflow met GenAI

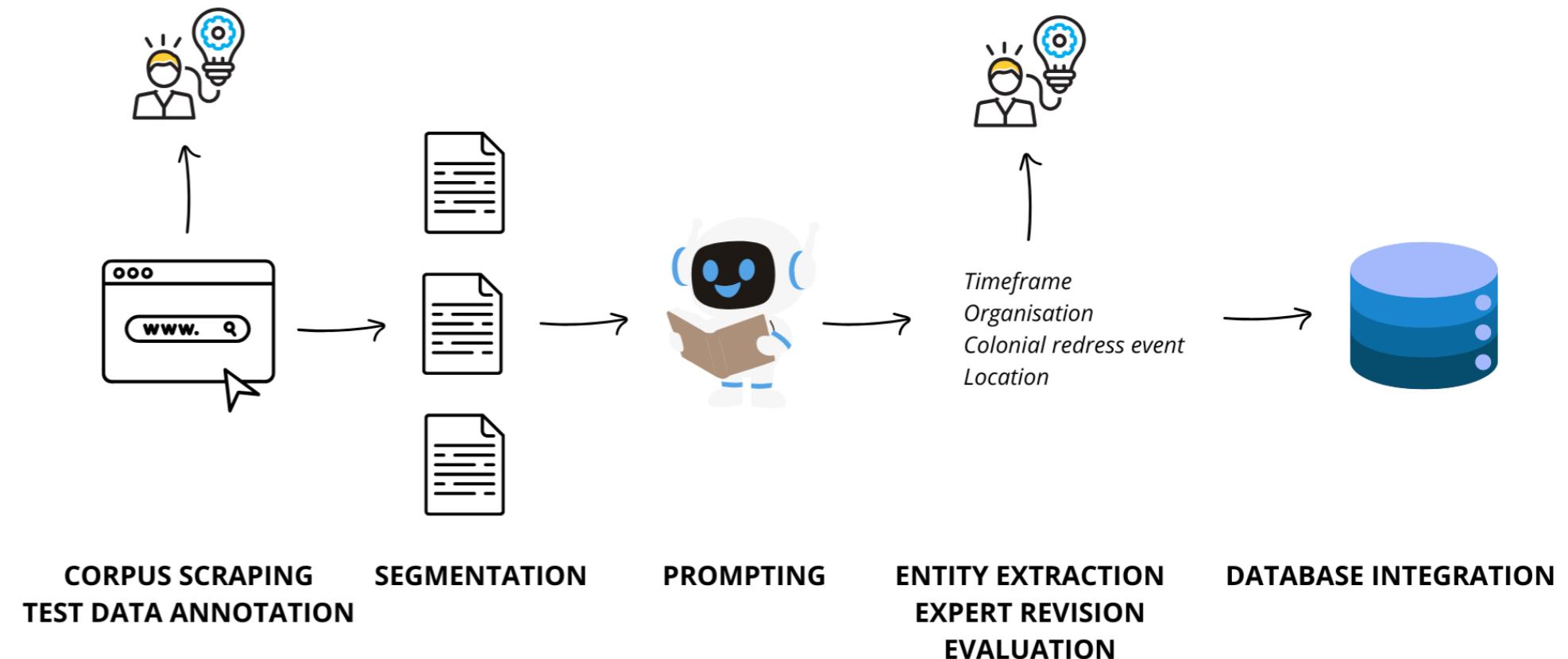
1. Data

- Bronnen manueel verzameld.
- Annotatie test-set voor kwantitatieve evaluatie semi-automatisch.

2. Entity-extraction met GenAI.

- Evaluatie: kwantitatief + HITL!

3. Database-integratie.



3. DESSERT

DESSERT CASE 1 – INDUCTIVE CONTENT ANALYSIS

BACKGROUND INFORMATION

- Case is part of research into how gender-neutral pronouns are interpreted and evaluated by readers
- Case data comes from online survey

Input given to participant	Piece of text from labour law of fictitious organisation on working from home (“ <i>Het personeelslid vermeldt steeds het tijdstip en de locatie van zijn afstandswerk in zijn professionele agenda. [...]</i> ”)
Versions of input	male pronouns – plural form – die-die-diens – die-hen-hun – hen-hen-hun
Output requested from participant	Envisage employee, give the employee a name, and write text of three sentences about the employee

EXAMPLE RESEARCH DATA

TEKST 1

Pieter De Backer is 35 jaar oud. Hij werkt in een bankkantoor en heeft 3 kinderen. Omdat hij alleenstaande papa is, wil hij graag wat meer afstandswerk kunnen doen van thuis uit.

TEKST 2

Elise werkt in een informatica bedrijf maar woont ver van het kantoor, waardoor ze veel tijd verliest in openbaar vervoer en ze zou efficienter willen zijn en thuis willen werken.

TEKST 3

Bart werkt voor een middelgrote onderneming die net de optie tot thuiswerk geïmplementeerd heeft. Hiervoor moet aan enkele voorwaarden voldaan worden. Dit reglement wordt duidelijk medegedeeld.

SPECIFIC GOAL CURRENT RESEARCH PHASE

- **Goal:** obtain exhaustive list of all possible gender representations based on participant responses (e.g., “male”, “female”, “transgender”, “ambiguous”, *none*, ...)
- **Type of research:** inductive qualitative research by means of content analysis of textual data

“TRADITIONAL” METHOD

- Researcher reads and analyses all responses
- Researcher defines taxonomy containing all possible categories, possibly informed by previous research

“GenAI-DRIVEN” METHOD (1)

- GenAI model defines taxonomy by means of **zero-shot prompting**

In the attached TXT file you will find twelve small Dutch texts that describe an employee of a fictitious company. Please provide an ****exhaustive**** overview of the genders assigned to the employee in the descriptions. Take into account perspectives (first-person versus third-person descriptions) and use of pronouns.

OUTPUT (1)

Totaaltelling (12 teksten)

- **Man:** 4 (Tekst 1, 3, 4, 9)  qualitativeResearch_in...
- **Vrouw:** 3 (Tekst 2, 5, 10)  qualitativeResearch_in...
- **Non-binair (of 'they/hen/hun' bedoeld):** 2 (Tekst 8, 12)  qualitativeResearch_in...
- **Ambigu/Ongespecificeerd:** 2 (Tekst 6, 7)  qualitativeResearch_in...
- **Niet van toepassing (geen individu):** 1 (Tekst 11)  qualitativeResearch_in...

Opmerkingen over perspectief en voornaamwoorden

- **Eerste persoon** (Teksten 4, 7, 10): hier wordt gender vooral afgeleid uit de **naam** van de spreker (Jan, Nancy) of is het **variabel** (ik/vriendin in Tekst 7). Voornaamwoorden zijn *ik/mijn*; gender komt niet uit het voornaamwoord zelf maar uit context/naam.  qualitativeResearch_in...
- **Derde persoon** (Teksten 1, 2, 3, 5, 6, 8, 9, 12): hier zijn **voornaamwoorden** en **rolwoorden** (bv. *hij, papa, ze, hen/hun*) de primaire signalen; waar die ontbreken, doet de **voornaam** het werk (Bart, Anna). Bij **unisex** namen (Robin, Chris) bepaalt de daadwerkelijke pronomengebruik (of het ontbreken daarvan) of de gender **ambiguous** blijft of **non-binair** wordt toegekend.  qualitativeResearch_in...
- **Onpersoonlijk** (Tekst 11): imperatieve zonder referentie aan een individu → geen genderassortiatie mogelijk.  qualitativeResearch_in...

“GenAI-DRIVEN” METHOD (2)

– Few-shot prompting

In the attached TXT file you will find twelve small Dutch texts that describe an employee of a fictitious company. Please provide an **exhaustive** overview of the genders assigned to the employee in the descriptions. Take into account perspectives (first-person versus third-person descriptions) and use of pronouns. Below you can find three examples that can help you define the gender labels.

EXAMPLE 1

- Text: Dirk is 55 jaar oud. Hij werkt in een advocatenkantoor en heeft 2 kinderen.

- Gender label: M name + M pronouns

EXAMPLE 2

- Text: De naam van het personeelslid is Katrien. Katrien is heel punctueel.

- Gender label: F name + no pronouns

EXAMPLE 3

- Text: Alex is net afgestudeerd en begonnen met werken voor dit bedrijf. Ze willen zich graag zo snel mogelijk integreren.

- Gender label: ambiguous name + plural pronouns

OUTPUT (2)

Tekst	Gender Label
1	M name + M pronouns
2	F name + F pronouns
3	M name + no pronouns
4	M name + first-person
5	F name + no pronouns
6	Ambiguous name + no pronouns
7	Mixed (speaker unspecified + partner F)
8	F name (current) + M name (past) + non-binary pronouns
9	M name + non-standard pronoun ("die")
10	F name + first-person
11	No name + no pronouns
12	Ambiguous name + plural pronouns

ANALYSIS GenAI USE

- Voordelen?
- Nadelen?
- Gevaren / valkuilen / ... ?

POTENTIAL QUESTIONS

- Wat is de beste werkwijze?
 - Eerst zelf (een beperkte subset van) de dataset lezen, verwerken en een taxonomie definiëren en daarna jouw taxonomie naast die van het GenAI-model leggen?
 - Eerst een GenAI model de data laten inlezen en een taxonomie definiëren, om die daarna als onderzoeker te fine-tunen?
- Welke aanpak is te verkiezen: zero-shot of few-shot?
 - Bij zero-shot geef je volledige “inductieve vrijheid” aan het model, maar kan de output misschien niet bruikbaar zijn
 - Bij few-shot stuur je de inductie in een bepaalde richting aan de hand van de voorbeelden die je meegeeft, maar heb je misschien meer garanties dat de output bruikbaar zal zijn

DESSERT CASE 2 – DEDUCTIVE CONTENT ANALYSIS

BACKGROUND INFORMATION

- Case is part of research into how gender-neutral pronouns are interpreted and evaluated by readers
- Case data comes from online survey

Input given to participant	Piece of text from labour law of fictitious organisation on working from home (“ <i>Het personeelslid vermeldt steeds het tijdstip en de locatie van zijn afstandswerk in zijn professionele agenda. [...]</i> ”)
Versions of input	male pronouns – plural form – die-die-diens – die-hen-hun – hen-hen-hun
Output requested from participant	Envisage employee, give the employee a name, and write text of three sentences about the employee

EXAMPLE RESEARCH DATA

TEKST 1

Pieter De Backer is 35 jaar oud. Hij werkt in een bankkantoor en heeft 3 kinderen. Omdat hij alleenstaande papa is, wil hij graag wat meer afstandswerk kunnen doen van thuis uit.

TEKST 2

Elise werkt in een informatica bedrijf maar woont ver van het kantoor, waardoor ze veel tijd verliest in openbaar vervoer en ze zou efficienter willen zijn en thuis willen werken.

TEKST 3

Bart werkt voor een middelgrote onderneming die net de optie tot thuiswerk geïmplementeerd heeft. Hiervoor moet aan enkele voorwaarden voldaan worden. Dit reglement wordt duidelijk medegedeeld.

SPECIFIC GOAL CURRENT RESEARCH PHASE

- Based on responses from participants, a taxonomy with all possible gender representations was built
- **Next step:** a second person analyses all responses and tries to assign them to one of the categories from the taxonomy to corroborate its validity
- **Type of research:** “secondary coder” for deductive qualitative research by means of content analysis of textual data

“TRADITIONAL” METHOD

- Second researcher performs secondary coding

“GenAI-DRIVEN” METHOD

- GenAI model performs secondary coding
- Prompt

In the attached TXT file you will find twelve small Dutch texts that describe an employee of a fictitious company. Please assign one of the following gender representation labels to the texts: M name + M pronouns, F name + F pronouns, M name + no pronouns, F name + no pronouns, ambiguous name + no pronouns, "Me and girlfriend", "Transgender", F name + 1PSPs, M name + 1PSPs. Note that "1PSPs" stands for "first-person singular pronouns". If you think none of the provided labels fit the text, please define a new label and clearly indicate this in the output.

OUTPUT (COPILOT, TAKE 1)

Tekst Nr.	Naam	Pronomen(s)	Label
1	Pieter De Backer	hij	M name + M pronouns
2	Elise	ze	F name + F pronouns
3	Bart	geen	M name + no pronoun:
4	Jan Leerdam	ik	M name + 1PSPs
5	Anna	geen	F name + no pronouns
6	Robin	geen	ambiguous name + no pronouns
7	geen	ik, mijn vriendin	Me and girlfriend
8	Kelly (voorheen Kevin)	hen/hun	Transgender
9	Henk Jan	die	M name + no pronoun:
10	Nancy	ik	F name + 1PSPs
11	geen	geen	New label: No name + no pronouns
12	Chris	ze	ambiguous name + no pronouns

OUTPUT (COPILOT, TAKE 2)

Text ID	Assigned Label
TEKST 1	M name + M pronouns
TEKST 2	F name + F pronouns
TEKST 3	M name + no pronouns
TEKST 4	M name + 1PSPs
TEKST 5	F name + no pronouns
TEKST 6	ambiguous name + no pronouns
TEKST 7	Me and girlfriend
TEKST 8	Transgender
TEKST 9	M name + M pronouns
TEKST 10	F name + 1PSPs
TEKST 11	No name + no pronouns (NEW LABEL)
TEKST 12	Ambiguous name + gender-neutral pronouns (NEW LABEL)



OUTPUT (ChatGPT)

Tekst	Label
1	M name + M pronouns
2	F name + F pronouns
3	M name + no pronouns
4	M name + 1PSPs
5	F name + no pronouns
6	ambiguous name + no pronouns
7	"Me and girlfriend"
8	Transgender
9	M name + no pronouns
10	F name + 1PSPs
11	NEW LABEL: No name + no pronouns
12	ambiguous name + no pronouns

ANALYSIS GenAI USE

- Voordelen?
- Nadelen?
- Gevaren / valkuilen / ... ?

POTENTIAL QUESTIONS (1)

- In welke mate “weet” het GenAI-model wat bedoeld wordt met de verschillende labels? Bijvoorbeeld dat “M” voor “masculine” staat en “F” voor “feminine”?
- Welke impact heeft prompting op de resultaten? Zijn de resultaten wel betrouwbaar als ze niet exact repliceerbaar zijn? (cf. Take 1 en Take 2 met Copilot)
- Wat als resultaten sterk verschillen bij een licht andere prompt? Of bij kiezen van ander model? Wat zegt dit over de betrouwbaarheid van de kwalitatieve codering?

POTENTIAL QUESTIONS (2)

- Om impact van uiteenlopende resultaten te minimaliseren: zelfde prompt vijf keer runnen met Model A en vijf keer met Model B en dan “majority vote” toepassen bij elk item uit de dataset? Valt dit nog steeds onder “verantwoord” en “duurzaam” gebruik?
- Kan GenAI deductief codeerwerk overnemen en effectief als een betrouwbare second coder ingezet worden? Welke minimale accuraatheid/overlap met menselijke codeerder zou hiervoor vereist zijn?
- Is hoge accuraatheid en consistentie in antwoorden verwachten van een AI-model wel terecht? Zijn wij als menselijke codeerders altijd 100% accuraat en consistent?

DESSERT CASE 3 –

INFERENTIAL

QUANTITATIVE ANALYSIS

BACKGROUND INFORMATION (1)

- Case is part of research into attitudes of L2 learners towards Intelligent Computer-Assisted Language Learning (ICALL)
- Participants: 23 L2 Spanish students
- Course: Spanish writing skills

BACKGROUND INFORMATION (2)

- Module: vocabulary learning (topic: business)
- Assignments in module: complete vocabulary learning exercises in online ICALL environment
- Students filled in a Likert questionnaire that gauges their attitudes towards ICALL **before** (“pre”) and **after** (“post”) completing the module

EXAMPLE RESEARCH DATA

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		Q1	Q2	Q3*	Q4	Q5	Q6	Q7*	Q8	Q9	Q10	Q11	Q12*	Q13*	Q14	Q15*
2	P1	7	4	1	7	6	6	4	7	7	6	6	2	6	7	3
3	P2	5	6	3	5	6	4	2	6	5	3	6	3	6	6	3
4	P3	5	2	4	3	6	6	6	7	7	5	7	6	7	7	1
5	P4	6	6	1	7	6	5	2	5	6	6	8	4	8	2	
6	P5	6	3	7	6	7	6	4	7	4	8	4	3	5	5	2
7	P6	6	4	6	4	4	6	4	5	5	4	6	4	7	6	1
8	P8	6	3	8	8	8	7	1	7	5	6	8	4	8	8	3
9	P9	5	7	6	4	6	7	2	5	5	5	6	3	7	7	8
10	P11	7	2	6	6	8	8	2	7	6	5	6	1	7	7	1
11	P12	8	6	8	7	8	6	8	5		5	3	5	4	7	1
12	P14	6	7	5	3	5	6	3	6	6	5	2	7	5	5	1
13	P16	5	6	6	5	4	5	4	2	2	3	4	6	6	6	1
14	P17	6	3	5	6	7	7	3	5	6	2	4	3	6	7	6
15	P18	5	6	4	5	6	7	2	5	5	6	5	1	5	6	3
16	P19	8	5	4	3	8	7	4	6	6	4	6	5	7	6	3
17	P20	6	6	5	6	6	6	3	7	6	7	6	3	7	7	4
18	P21	6	6	3	5	7	5	3	5	6	6	6	2	6	7	8
19	P22	6	4	7	4	6	4	4	4	5	2	6	7	7	6	2
20	P23	4	7	4	6	5	6	5	7	5	7	7	2	7	7	7
21	P24	8	8	6	7	7	7	1	7	7	5	8	1	2	8	1
22	P25	5	8	6	4	6	6	2	7	5	4	7	2	7	8	1
23	P27	6	1	4	6	7	5	5	6	4	5	7	6	5	6	1
24	P28	6	8	3	5	7	7	6	6	5	4	7	7	5	7	3
25	avg	6.00	5.13	4.87	5.30	6.35	6.04	3.48	5.83	5.36	4.91	5.87	3.78	6.00	6.70	2.87
26	stdev	1.04	2.07	1.91	1.43	1.15	1.02	1.73	1.27	1.14	1.53	1.58	2.00	1.35	0.88	2.28

SPECIFIC GOAL CURRENT RESEARCH PHASE

- **Goal:** obtain quantification of (potential) difference between pre- and post-scores → Which change in attitude? Is this change statistically significant or not?
- **Type of research:** inferential quantitative research

“TRADITIONAL” METHOD

- Researchers familiarise themselves with the topic and look up different possibilities of performing the analysis
 - Website (trustworthy institutions!)
 - Papers
 - Documentation of statistical analysis tools (SPSS, SciPy module in Python, ...)
- Consult an expert to corroborate findings

Statistical tests: Continuous data

<https://www.brookes.ac.uk/students/academic-development/mathematics-and-stats/statistics/inferential-statistics/continuous-data>

This page contains general information for choosing commonly used statistical tests. The examples linked provide general guidance which should be used alongside the conventions of your subject area. Where possible, a brief explanation of the test is given with links to performing this test using Excel, SPSS and R. It is worth noting that the examples often contain information about interpreting the output and results so can act as a guide to interpreting statistical results too.

To navigate this table, consider the following questions:

- Is your outcome variable continuous?
- Does the data meet the requirements for a parametric test (e.g. normality)? Notice, that for each parametric test, where possible, a corresponding nonparametric test is presented.
- How many samples (or groups) do you have?
- Are the outcomes paired (or dependent)?

Parametric		Nonparametric	
1 Sample	One sample t test If sigma is unknown, use a one sample t test to determine if the sample is likely to have come from a given population with a defined mean.	Example	Excel
		SPSS	R
2 Samples Paired (or dependent, repeated measures)	Paired t test The paired-samples t test is used when the data is from related, paired or longitudinal samples.	Example	Excel
		SPSS	R
2 Samples Unpaired (or independent)	Unpaired t test. An unpaired t test is used to assess if the mean values of two independent samples are equal. Firstly, you need to assess equality of variances using an F-test, details of which are given within the examples below.	Example	Excel ($\sigma_1=\sigma_2$)
			SPSS
		Excel ($\sigma_1 \neq \sigma_2$)	R
1 Sample	One sample sign (or median) test A one sample sign test is used to explore if the median of the sample data is equal to a given value.	Example	Excel
		SPSS	R
2 Samples Paired (or dependent)	Wilcoxon signed rank test The Wilcoxon signed ranks test is used to compare the medians of two related samples.	Example	Excel
		SPSS	R
2 Samples Unpaired (or independent)	Mann-Whitney U test The Mann-Whitney U test is used to compare the medians of two independent samples.	Example	Excel
		SPSS	R

Dependent T-Test using SPSS Statistics

Introduction

The dependent t-test (called the paired-samples t-test in SPSS Statistics) compares the means between two related groups on the same continuous, dependent variable. For example, you could use a dependent t-test to understand whether there was a difference in smokers' daily cigarette consumption before and after a 6 week hypnotherapy programme (i.e., your dependent variable would be "daily cigarette consumption", and your two related groups would be the cigarette consumption values "before" and "after" the hypnotherapy programme). If your dependent variable is dichotomous, you should instead use [McNemar's test](#).

This "quick start" guide shows you how to carry out a dependent t-test using SPSS Statistics, as well as interpret and report the results from this test. However, before we introduce you to this procedure, you need to understand the different assumptions that your data must meet in order for a dependent t-test to give you a valid result. We discuss these assumptions next.

Wilcoxon Signed-Rank Test using SPSS Statistics

Introduction

The Wilcoxon signed-rank test is the nonparametric test equivalent to the [dependent t-test](#). As the Wilcoxon signed-rank test does not assume normality in the data, it can be used when this assumption has been violated and the use of the dependent t-test is inappropriate. It is used to compare two sets of scores that come from the same participants. This can occur when we wish to investigate any change in scores from one time point to another, or when individuals are subjected to more than one condition.

For example, you could use a Wilcoxon signed-rank test to understand whether there was a difference in smokers' daily cigarette consumption before and after a 6 week hypnotherapy programme (i.e., your dependent variable would be "daily cigarette consumption", and your two related groups would be the cigarette consumption values "before" and "after" the hypnotherapy programme). You could also use a Wilcoxon signed-rank test to understand whether there was a difference in reaction times under two different lighting conditions (i.e., your dependent variable would be "reaction time", measured in milliseconds, and your two related groups would be reaction times in a room using "blue light" versus "red light").

This "quick start" guide shows you how to carry out a Wilcoxon signed-rank test using SPSS Statistics, as well as interpret and report the results from this test. However, before we introduce you to this procedure, you need to understand the different assumptions that your data must meet in order for a Wilcoxon signed-rank test to give you a valid result. We discuss these assumptions next.

<https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php>

```
scipy.stats.
```

wilcoxon

```
wilcoxon(x, y=None, zero_method='wilcox', correction=False,  
alternative='two-sided', method='auto', *, axis=0, nan_policy='propagate',  
keepdims=False) \[source\]
```

Calculate the Wilcoxon signed-rank test.

The Wilcoxon signed-rank test tests the null hypothesis that two related paired samples come from the same distribution. In particular, it tests whether the distribution of the differences $x - y$ is symmetric about zero. It is a non-parametric version of the paired T-test.

“GenAI-DRIVEN” METHOD

– Input prompt

Have a look at the attached spreadsheet. The sheet `results_pre` contains the responses on a pre-test questionnaire; the sheet `results_post` contains the post-test results. I want to apply a statistical test to these data to find out whether there is a statistically significant difference between the pre- and post-results. Which test do I need?

OUTPUT (COPILOT – 1)

You've got the **same people measured twice** (pre-test vs post-test) on a Likert-type questionnaire (Q1–Q15; items with asterisks are reverse-worded so that "higher = better"). That design calls for a **within-subjects (paired) comparison**.  [quantitativeAnalysis_i...](#)

Which test?

Use a **paired-samples t-test** on a **composite score** (e.g., the participant's mean across Q1–Q15), provided you're comfortable treating the composite as approximately interval and the **within-person differences are roughly normal**. With 15 items aggregated and your sample size, this is the common and appropriate choice.

If you prefer not to make that assumption (or if the differences look clearly non-normal), use the **Wilcoxon signed-rank test** on the same composite. Both tests address the *same* research question ("Did the average score change from pre to post?"), but the Wilcoxon test is non-parametric.

OUTPUT (COPILOT – 2)

Recommendation: which test?

- **Primary test:** Paired-samples t-test on the composite (overall mean across items).
 - Rationale: within-subjects design, composite scale with good reliability, and paired differences approximately normal.  quantitativeAnalysis_i...
- **Non-parametric alternative** (if you prefer not to rely on normality or to be extra conservative): Wilcoxon signed-rank test on the same composite.  quantitativeAnalysis_i...

In short: for your data as provided, the paired t-test on the composite is the appropriate choice.

OUTPUT (ChatGPT – 1)

Good question — to choose the right statistical test, we need to consider **the design of your data** and the **scale of measurement** of your variables.

Here's how to decide:

- ◆ **Step 1: Check the study design**

Since your file has:

- `results_pre` : pre-test scores
- `results_post` : post-test scores

...it's **the same participants tested twice** (a *within-subjects / repeated measures design*).

That means we're comparing **paired** measurements.

- ◆ **Step 2: Check the data type**

Open the spreadsheet and look at the dependent variable (e.g., scores, Likert-scale responses, etc.):

- If the responses are **interval or ratio scale** (e.g., numeric scores, test percentages, continuous ratings):
→ use a **paired samples t-test** (parametric).

- Assumptions:

- Data are approximately normally distributed.
- Differences between pre and post are continuous and symmetric.
- If the responses are **ordinal** (e.g., Likert 1–5, 1–7) or **not normally distributed**:
→ use a **Wilcoxon signed-rank test** (non-parametric alternative).



OUTPUT (ChatGPT – 2)

Inspect the dataset and determine which test I need

Great — here's what we can tell from the inspection:

- Both sheets (`results_pre` and `results_post`) have **25 participants (P1–P25)**.
- There are **15 questions (Q1–Q15)**, mostly numeric (floating-point values).
- Responses appear to be **integer-scale ratings (1–8 range)** → likely **Likert-type items** (ordinal, treated as approximately continuous).
- There are only a few missing values (e.g., one missing in Q9, Q13).

That means:

- We have *paired* data (same participants, before and after).
- Each column is a repeated measure of the same question.

To decide between a **paired t-test** and a **Wilcoxon signed-rank test**, we'll check whether the **differences between pre and post scores are normally distributed** across participants.

Let's test that next. 

Here's what the inspection shows:

- The **average difference** between post- and pre-test means is **-0.19**, meaning there's a slight *decrease* on average.
- The **Shapiro–Wilk normality test** gives **p = 0.12**, which is **greater than 0.05**, so the differences are **approximately normally distributed**.

 **Conclusion:**

Because the data come from the same participants (paired) and the difference scores are normally distributed,
the appropriate test is a **paired samples t-test** (parametric).



OUTPUT (Le Chat – 1)

 To determine the appropriate statistical test for comparing your pre-test and post-test results, we need to consider the following:

Key Considerations

1. **Paired Data:** Since you are comparing the same participants before and after an intervention, your data is paired.
2. **Data Type:** The nature of your data (e.g., continuous, ordinal, or categorical) will influence the choice of test.
3. **Normality:** Whether the differences between the paired data are normally distributed.

Common Tests for Paired Data

- **Paired t-test:** Used when the differences between paired data are normally distributed and the data is continuous.
- **Wilcoxon Signed-Rank Test:** A non-parametric alternative to the paired t-test, used when the differences are not normally distributed or the data is ordinal.

Next Steps

1. **Check the Data Type:** Confirm if your data is continuous, ordinal, or categorical.
2. **Check for Normality:** If the data is continuous, check if the differences between the pre- and post-test results are normally distributed.

I will now load the spreadsheet and inspect the data in the `results_pre` and `results_post` sheets to help determine the most appropriate test.

OUTPUT (Le Chat – 2)

Recommendation for Statistical Test

- **Paired t-test:** Suitable for questions where differences are normally distributed.
- **Wilcoxon Signed-Rank Test:** More appropriate for questions where differences are not normally distributed.

Given the mixed results, the **Wilcoxon Signed-Rank Test** is a safer choice overall, as it does not assume normality and is suitable for ordinal or non-normally distributed data.

Would you like me to proceed with the Wilcoxon Signed-Rank Test for all questions, or do you have a preference for specific questions?

ANALYSIS GenAI USE

- Voordelen?
- Nadelen?
- Gevaren / valkuilen / ... ?

POTENTIAL QUESTIONS (1)

- Kan GenAI als een extra bron van “inleesmateriaal” dienen?
- Is de informatie inhoudelijk correct? Is de informatie helder beschreven?
- Verschillende AI-modellen combineren om te zien of hun output overeenkomt en, indien ja, ervan uitgaan dat de suggestie correct is?
- Kan GenAI functie van expert vervullen die op basis van je inleeswerk de finale suggestie doet?

POTENTIAL QUESTIONS (2)

- Wat met tegenstrijdigheden?
 - ChatGPT geeft aan dat Likertdata Wilcoxon-test vereist, maar kiest uiteindelijk voor gepaarde *t*-test
 - Le Chat stelt Wilcoxon-test voor terwijl Copilot en ChatGPT voor gepaarde *t*-test kiezen
- Wat met het aanbod om effectieve berekening van de test ook toe te vertrouwen aan het AI-model? (zie screenshots volgende slide)

If you want, I can run it for you

I can:

- build an overall composite (and optional subscales),
- check normality of the paired differences,
- run the appropriate paired test(s),
- report the **effect size** (e.g., Cohen's *dz*) and **95% CI**, and
- optionally do **item-level** analyses with a multiple-comparisons adjustment.

Would you like me to (a) test the **overall composite**, (b) **each item** separately, or (c) **define subscales** and test those? If you have a scoring key for subscales, send it over and I'll use that.

If you'd like, I can:

- Inspect your dataset (`results_pre` and `results_post`),
- Check whether the data meet normality assumptions, and
- Run the correct test for you (t-test or Wilcoxon).

Would you like me to do that?

FINALLY ... THE END!



Joanna Maciejewska
@AuthorJMac

...

You know what the biggest problem with pushing all-things-AI is? Wrong direction.

I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes.

12:50 PM · Mar 29, 2024 · 3.2M Views

Nele Kellens
Miriam Taverniers
Janna Everaert
Jasper Degraeuwe

FACULTY OF ARTS AND PHILOSOPHY

www.ugent.be

-  Universiteit Gent
-  @ugent
-  @ugent
-  Ghent University