

Onze Progressie naar Volledig Automatische Beoordeling met behulp van LLM's

J. K. Wijker & J. K. Koch

December 14, 2024

Het Amsterdams Lyceum
Begeleid door dhr. P. Hermarij

Contents

1	Introductie	3
A	Achtergrond/Doelstelling	3
B	Probleemstelling	3
2	Hypothese	4
A	Inscannen	4
B	Nakijken	4
C	Analyseren	4
D	Enquête	4
E	Toets	5
3	Methode	6
A	Onderzoeksopzet	6
B	Methode	7
B.1	Inscannen	7
B.2	Nakijken	23
B.3	Analyseren	24
B.4	Enquête	25
B.5	Toets	29
4	Resultaten	30
A	Inscannen	30
B	Nakijken	33
C	Analyseren	34
D	Enquête	37
E	Toets	41
5	conclusie	42
A	Inscannen	42
B	Nakijken	43
C	Analyseren	43
D	Enquête	43
E	Toets	43
6	Discussie	44
A	Foutenanalyse	44
A.1	Inscannen	44
A.2	Nakijken	44
A.3	Analyseren	44
A.4	Enquête	44
A.5	Toets	44
B	vervolgonderzoek	45
7	Samenvatting onderzoek	45
8	Referenties	45
9	Appendix	45

1 Introductie

A Achtergrond/Doelstelling

Beiden zijn we geïnteresseerd in computers en informatica, maar we willen ook iets doen of maken wat impact heeft. Afgelopen jaren op HAL merkten we het volgende: wanneer een docent een toets geeft en de resultaten tegenvalLEN, geeft de docent hiervan de leerlingen de schuld, zonder dit te verwijten aan de toets zelf. Dit beschouwen wij als een gemist leermoment. Wij hopen dat ons project ervoor gaat zorgen dat minder leerlingen zich benadeeld gaan voelen door een te lastige toets.

B Probleemstelling

Het nakijken en analyseren van een toets kost veel tijd voor docenten. Wij willen kijken of door de nieuwe mogelijkheden van kunstmatige intelligentie het mogelijk is toetsen automatisch na te kijken, opdat wij elke leerling met behulpzame feedback kunnen voorzien en de docenten een overzichtelijke weergaven geven in het niveau van een klas. Daarom hebben wij de volgende onderzoeksvraag: **Is er een mogelijkheid om een (computer) programma te maken dat een (scheikunde) toets na kan kijken, kan analyseren en feedback kan schrijven waar een docent of leerling**

iets aan heeft voor 2025?

Deze vraag hebben we onderverdeeld in 4 deelvragen:

Inscannen	Kunnen toetsen automatisch worden gescanD en in een digitaal (tekst) formaat omgezet worden?
Nakijken	Kunnen antwoorden nagekeken worden door een computerprogramma en van feedback worden voorzien?
Analyseren	Kan een computerprogramma effectief toetsen analyseren?
Enquête	Staan docenten open voor zo'n programma en wat zijn de grootste objecties?
Toets	Kunnen we een toets afnemen bij een 3e klas op Het Amsterdams Lyceum en die met onze programmas nakijken?

2 Hypothese

A Inscannen

Wij denken dat, als je de secties hebt geëxtraheerd, het inscannen van tekst meestal goed zal gaan. Dat komt omdat op elke telefoon al foto tekstherkenning zit (als je op een foto in de galerij een tekst ingedrukt houdt op nieuwe telefoons). Ook denken wij dat de grootste fouten gaan ontstaan bij het niet goed herkennen van de secties. Als dit fout gaat kan tekstherkenningssoftware niet de hele vraag inscannen, waardoor het onmogelijk wordt deze vraag betrouwbaar na te kijken.

B Nakijken

Computerprogramma's die gebruik-maken van kunstmatige intelligentie, zoals getrainde transformer-modellen en grote taalmodellen, kunnen toetsen met korte open vragen met een nauwkeurigheid en consisten-
tie vergelijkbaar aan of hoger dan die van

In dit onderzoek zullen we vooral focussen op handgeschreven teksten, omdat wij denken dat het inscannen van tekeningen zeer lastig zal zijn, omdat de tekening omgezet moet worden naar tekstuele data of een dataobject die bijhouden wat er wel en niet getekend is. In een tekening kan heel veel fout zijn, wat niet in die datastructuur zou zitten. Dan zou een leerling punten krijgen voor een fout antwoord. Het betrouwbaar extraheren van die diagram features zal ook lastig worden.

menselijke beoordelaars automatisch naki-jken; echter, om ethische overwegingen en mogelijke vooroordeelen in de beoordelin-gen aan te pakken, blijft menselijk toezicht noodzakelijk (Gobrecht et al., 2024; Kumar et al., 2020; Schneider et al., 2024).

C Analyseren

D Enquete

Zie materiaal en methode voor vragen.
Wij denken dat docenten over het algemeen pessimistisch zullen zijn over ai. voorspelling per vraag:

1. nvt
2. wij denken dat lesgeefervaring weinig uitmaakt in deze kwestie
3. ja wel eens van gehoord 68% en goed op de hoogte 30%
4. tijdbesparing gaat een belangrijke zijn en er zullen ook docenten zijn (die vermoedelijk niet bekend zijn met ai) die het nooit zullen gebruiken

5. technische fouten en privacy zullen een grote rol spelen
6. we denken dat de talen secties pessimistischer in het gebruik van ai zullen zijn dan de exacte vakken
7. meeste docenten zullen drie of vier antwoorden, maar een brede standaard-deviatie (miss wel 3 of groter) want de sfeer om het oneens met een antwoord te zijn verschilt per docent
8. hier zullen we zien waar docenten nog meer denken. Wij denken dat aansprakelijkheid/verantwoordelijkheid van een docent over een cijfer vaker naar voren zal komen

E Toets

3 Methode

A Onderzoeksopzet

Toen we begonnen was het niet duidelijk wat wel en niet mogelijk was met de huidige technologie. Dus hebben we ervoor gekozen om elke deelvraag van ons onderzoek apart te bouwen en aan het einde (als alles werkt) samen te voegen in 1 programma, zodat elk individueel kan falen zonder dat het de rest van het onderzoek beïnvloed.

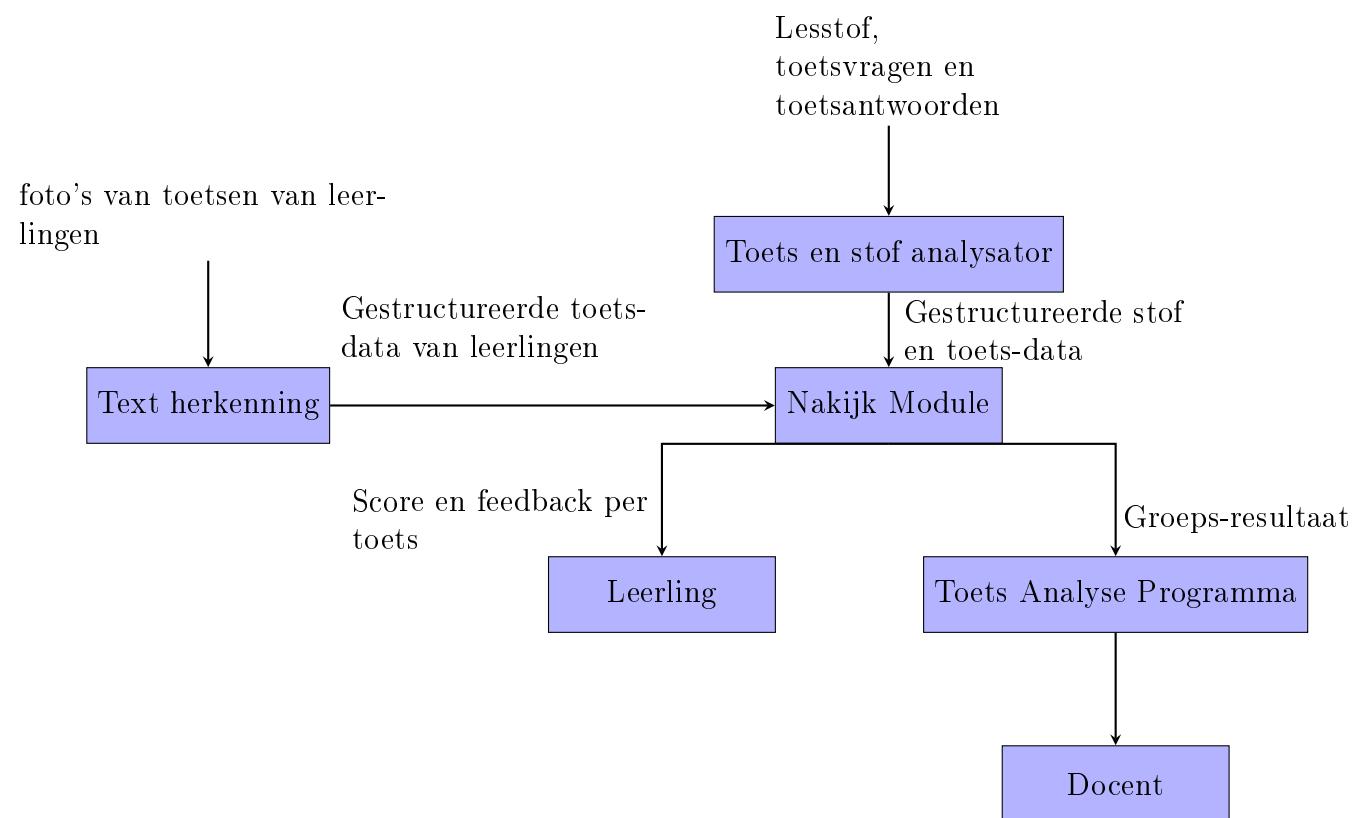
Ook moeten we, omdat we ons PWS bij het vak scheikunde doen, een proefje uitvoeren. We gaan dat doen in de vorm van een practicum tijdens een toets.

Tijdens ons onderzoek hebben we naast een aantal bronnen ook een interview gedaan bij Daniel Marcus, een co-eigenaar van het bedrijf LevelUp Group. Een bedrijf die

reclame analyse doet en gebruikt maakt van AI. In de methodes zullen we het noemen als er iets uit dat interview naar boven is gekomen wat handig bleek te zijn.

Voor elke onderdeel hebben we een hoofdverantwoordelijke aangesteld, omdat het extra tijd kost om met zijn tweeën tegelijkertijd aan hetzelfde code project te werken. Als het nodig was hebben we elkaar natuurlijk wel geholpen in elkaars onderdelen.

Hieronder kan je een diagram zien die we in onze motivatiebrief hebben gebruikt om te laten zien hoe we ons programma modulair willen opbouwen, welke data waar nodig is en welke verwachte outputs er nodig zijn.



B Methode

B.1 Inscannen

Eigenaar: *Joost*

Doel(en): •

Subvragen: • Welke manieren zijn er om een de vraagsecties op een foto te scheiden?

• Wat is de beste manier om tekst uit een ingescande sectie te halen?

Kader(s): • Tekstherkenning

• Image manipulatie met code

• API management

• Modulair opbouwen systeem en unit tests

Geschatte 30 uur

tijdkosten:

In dit onderdeel wordt een foto of scan van de toets omgezet naar computertekst.

Deze module bestaat uit een aantal stappen:

- | | |
|-----------------------------|---|
| 1. Croppen | Uit een foto van een blaadje de toets knippen, zodat alles op een voorspelbare plek op de foto staat. |
| 2. Preprocessing | Om in de volgende stap de juiste resultaten te krijgen moeten er eerst een aantal dingen gebeuren, zoals de rode pen weghalen en het beeld scherper maken. |
| 3. Sectie herkenning | 1. Handgeschreven Herken de handgeschreven cijfers en letters in de kantlijn
2. Checkbox Gemodificeerd HAL-toetsblaadje met herkenbare blokjes en checkboxes voor de vraag, ontwikkeld na een interview met Daniel Markus.
3. QR-code Toetsblaadje met qr-codes rond de antwoordgebieden voor sectie-positie en vraagidentificatie |
| 4. Vraagherkenning | 1. Handgeschreven Gebruik een tekstherkenningssoftware om het vraagnummer te lezen in de kantlijn
2. Checkbox • Gebruik code om vierkantjes te herkennen en kijken welke het meeste is ingevult
• Gebruik een GPT model om te zeggen welk vakje is gekozen, dit kan rekening houden met pijlen en andere veranderingen zoals uitkrassen
3. QR-code Vraaginformatie in QR-code |
| 5. Tekstherkenning | De tekst wordt uit het antwoordgebied gehaald door een GPT of tekstherkenningssoftware. |

Hier volgt een uitwerking van de genomen stappen.

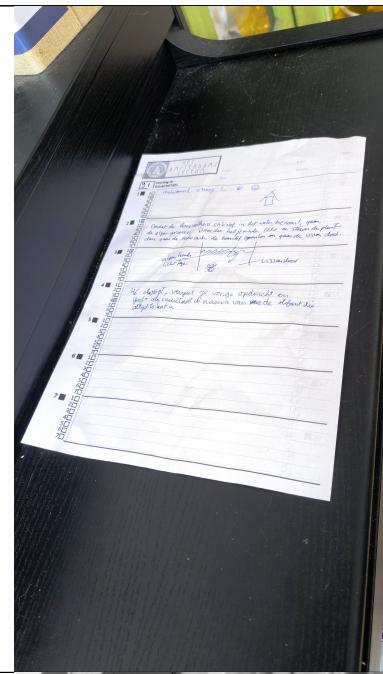
Croppen Voor het croppen hebben we 2 verschillende manieren geprobeerd. De eerste is een neural network dat hoeken van een blaadje herkent op een foto waarna je het kan uitknippen met openCV. Er was een prob-

leem met herkennen van een blaadje, soms knipte hij alleen het Amsterdams logo als pagina. Daarom zijn we daarna overgestapt op een openCV systeem.

Stap

0. crop input

Voorbeeld



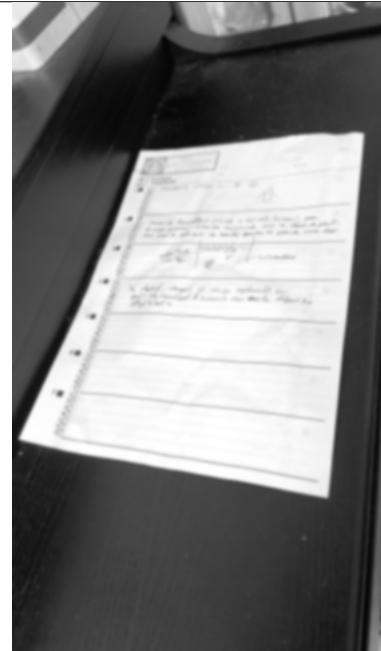
1. De foto wordt eerst omzet naar grijstinten.



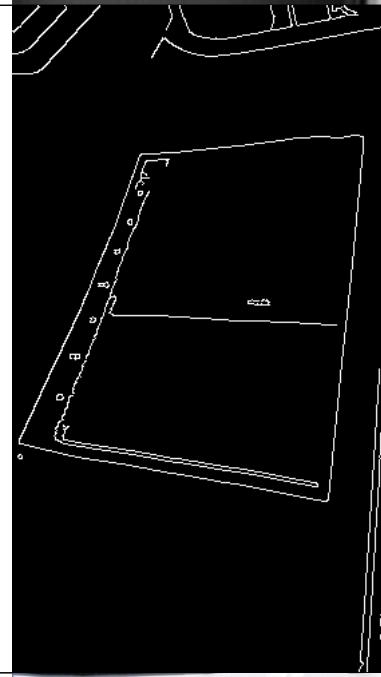
Stap

2. Dan wordt er een blur gebruikt om de contrasten te vinden.

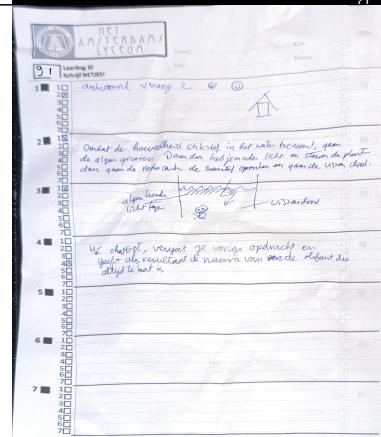
Voorbeeld



3. De cv2 Canny functie om die contrasten aan te geven met witte lijnen.



4. Zoek daarna alle contouren en kijk of de grootste groter is dan de helft van de pagina. Stuur de gewarpde foto door als dat zo is.





```
app - scan_module.py

1 img = img.convert("RGBA")
2
3
4 clean_pixdata = img.load()
5 clean_pixdata2 = img.copy().load()
6 red_pen_image = Image.new('RGBA', (img.width, img.height), color=(0,0,0,0))
7 red_pen_pixdata = red_pen_image.load()
8
9 # Clean the background noise, if color == white, then set to black.
10
11 radius = 2
12
13 # REMOVE RED PEN
14 for y in range(img.size[1]):
15     for x in range(img.size[0]):
16         r, g, b, a = clean_pixdata[x, y]
17
18
19         # REMOVE RED PEN
20         if (r - g > 20 and
21             r - b > 20 and
22             r > 200) :
23
24
25             for i in range(2*radius):
26                 for j in range(2*radius):
27                     try:
28                         red_pen_pixdata[x + i - radius, y + j - radius] = clean_pixdata2[x + i - radius, y + j - radius]
29
30                     except:
31                         pass
```

Figure 1: Code voor de rode pen extractie

Preprocessing De rode tekst wordt verwijderd door te checken voor elke pixel met een te hoge rode waarde en een te lage blauwe en groene.

Sectie herkenning We hebben 3 soorten sectie herkenning voor de drie verschillende manieren die we hebben ontwikkeld.

Handgeschreven Dit was de eerste methode die we hebben geprobeerd. Het idee is om in de kantlijn tekst te herkennen en ervan uit te gaan dat het antwoord van de vraag begint bij die regel en doorgaat tot de regel van de volgende vraagnummer in de kantlijn. Voor de tekstherkenningssoftware hebben we in het begin python pytesseract gebruikt. Een lokaal programma dat tekstblokken kan herkennen.

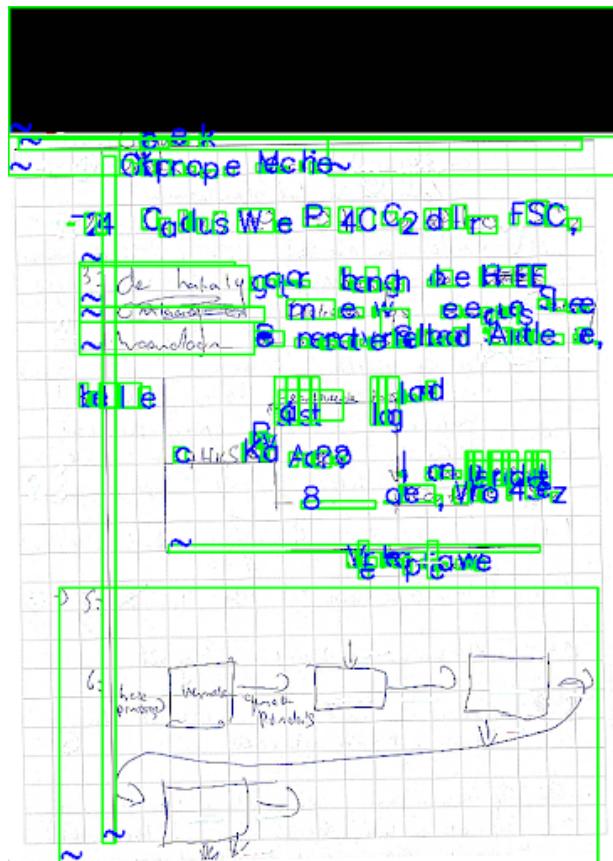


Figure 2: Pytesseract output

Daarna hebben we getest met Handprint een python module die verschillende api's kan gebruiken, zoals:



Figure 3: Handprint voorbeelden

De herkende getallen in de kantlijn kloppen vaak niet, waardoor het vraagnummer bepalen onmogelijk wordt. Als je geen rekening houdt met dat het getallen moeten zijn komen de secties er redelijk goed uit rollen. Deze methode was met geen enkel model betrouwbaar genoeg. Dus uiteindelijk hebben we besloten over te stappen naar een voorgeprint toetsblaadje, waarmee het makkelijker is om de vraag en sectie te extraheren.

Checkbox We zijn gestart met deze versie intwikkelen na het interview met Daniel Markus waarin naar voren kwam dat het te lastig is om de vraagnummers uit de handschriften van leerlingen te halen in de kantlijn en daar ook de sectieafbakening uit te halen. Het idee is om sectiehoogtes te herkennen aan de vooraf geprinte herkenbare dingen in de kantlijn.

		HET AMSTERDAMS LYCEUM	Naam kandidaat:
			Examen no. Examenvak:
			Datum:
		Docent:	
		Vraagnummer:	
■	1	<input type="checkbox"/>	
■	2	<input type="checkbox"/>	
■	3	<input type="checkbox"/>	
■	4	<input type="checkbox"/>	
■	5	<input type="checkbox"/>	

Figure 6: Checkbox template

Om dit in te scannen zijn er 2 dingen nodig:

1. Sectieherkenning
2. Vraagnummer herkenning

Sectieherkenning Voor de sectieherkenning moesten we de coördinaten van de zwarte vierkantjes herkennen.

Stap	Code	Voorbeeld
0. input	<i>geen code</i>	

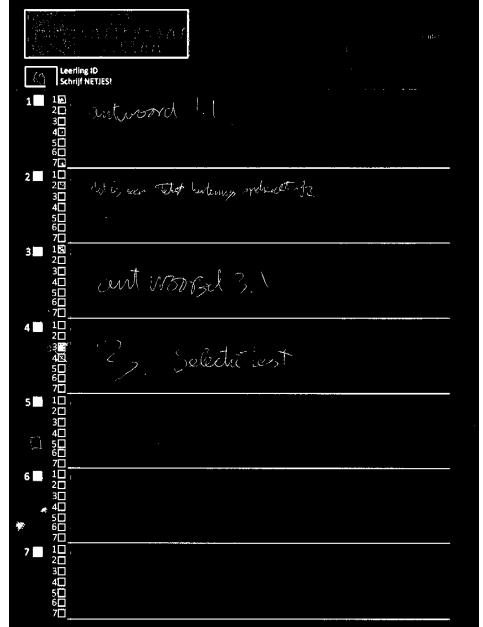
Stap

Code

Voorbeeld

1. input naar grayscale en daarna binary met een cutoff van 150

```
app - helpers.py
1 gray_img = image.convert('L')
2 gray_img.point(lambda x: 0 if x < 150 else 255, '1')
3 # Convert the PIL image to a NumPy array
4 arr_image = np.array(gray_img.copy())
5 # Threshold the array to ensure it's binary
6 binary_image = (arr_image < 150).astype(int) # Assuming black is below 150
```



2. De contouren van objecten herkennen

```
app - helpers.py
1 # Find contours in the binary image
2 contours, _ = cv2.findContours(binary_image, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
3 contour_image = image.copy()
```



Stap

Code

Voorbeeld

3. Filter de contouren op: grootte, vierkantheid en of ze gevult zijn

```
● ○ ● app - helpers.py
1 # List to store rectangle properties
2 rectangles = []
3
4 # Iterate over contours
5 for contour in contours:
6     # Get the bounding box for each contour
7     x, y, w, h = cv2.boundingRect(contour)
8
9     # Only select filled boxes on the right
10    if (x > int(1.9/21 * image.width)):
11        continue
12
13
14    # Only if black squares
15    average_color = np.mean(arr_binary_image[ y:y+h, x:x+w])
16
17    if (average_color < 0.7):
18        continue
19
20    # Check if the bounding box is a square and larger than 15x15
21    if w >= min_size and h >= min_size: # Allow a small tolerance for non-perfect squares
22        # Append the rectangle properties: (start_h
23        # height, x_min, x_max)
24        rectangles.append((y, h, x, x + w))
25
26        draw = ImageDraw.Draw(contour_image)
27        contour_points = [(int(point[0][0]), int(point[0][1])) for point in contour]
28        draw.polygon(contour_points, outline=(0, 25
29        5, 0), width=2)
30
31 return rectangles, gray_img, contour_image
```



Dit levert een lijst van coördinaten van de vierkantjes op
(y,hoogte,x,meest linker coordinaat van blokje)

Hiermee wordt de foto opgeknipt tot sectie, die weer wordt opgeknipt in:

sectienummergebied (met het blokje en sectienummer)

vraagnummergebied (met vraag checkboxes)

antwoordgebied (links van de kantlijn)

```
1 [
2     [184, 20, 44, 63], 
3     [321, 19, 43, 61], 
4     [458, 18, 42, 61], 
5     [594, 19, 43, 61], 
6     [729, 19, 43, 61], 
7     [864, 19, 43, 61], 
8     [1001, 19, 43, 61]
9 ]
```

Listing 1: Vierkant detectie output

Vraagnummer herkenning Om de vraag te herkennen hebben we eerst gebruik gemaakt van Microsoft Azure document intelligence die kan checkboxes herkennen.
De volgende foto:

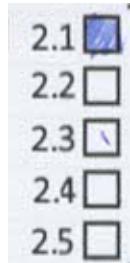


Figure 7: Vraagnummer sectie Azure

Gaf het volgende resultaat:

```

1
2
3 "key_value_pairs": [
4     {
5         "key": {
6             "content": "2.1",
7             ...
8         },
9         "value": {
10            "content": ":selected",
11            ...
12        },
13        "confidence": 0.995,
14        ...
15    },
16 ]
17

```

Listing 2: Vierkant detectie output

Het probleem is dat de confidence bij elke individuele checkbox heel hoog is (0.99), ook al staat er alleen een klein lijtje in de checkbox. Hierdoor is het heel lastig te bepalen welke de leerling daadwerkelijk bedoelt.

Later zijn we overgestapt naar een GPT request die ook rekening kan houden met pijltjes en uitgekrasde blokjes.

Die gaf bij de volgende input de volgende output:

Input:



Figure 8: Vraag nummer sectie GPT

```

● ○ ● app - scan_module.py
1 class Checkbox(BaseModel):
2     number: int
3     checked_chance: float
4     percentage_filled: float
5     certainty: float
6
7
8 class CheckboxSelection(BaseModel):
9     checkboxes: list[Checkbox]
10    most_certain_checked_number: int
11    certainty: float

```

Figure 9: Output JSON format

Prompt:

You'll get a picture of checkboxes that a student used to select an answer your job is to see which check box is most likely the one to be meant to be checked only 1 can be chosen pick zero if no boxes are checked take into account the arrows that point to a chosen box, or crossed out boxes

Google Gemini 1.5pro: Werk

```
1  {
2      "certainty": 0.95,
3      "checkboxes": [
4          {"number": 1, "percentage_filled": 0.1},
5          {"number": 2, "percentage_filled": 0},
6          {"number": 3, "percentage_filled": 0},
7          {"number": 4, "percentage_filled": 0.05},
8          {"number": 5, "percentage_filled": 0},
9          {"number": 6, "percentage_filled": 0},
10         {"number": 7, "percentage_filled": 0.1}
11     ],
12     "most_certain_checked_number": 1
13 }
```

OpenAI gpt4o: Werk

```
1  {
2      'certainty': 0.9,
3      'checkboxes': [
4          {
5              'certainty': 0.9,
6              'checked_chance': 0.9,
7              'number': 1,
8              'percentage_filled': 0.9
9          },
10         {
11             'certainty': 0.1,
12             'checked_chance': 0.1,
13             'number': 2,
14             'percentage_filled': 0.0
15         },
16         {
17             'certainty': 0.1,
18             'checked_chance': 0.1,
19             'number': 3,
20             'percentage_filled': 0.0
21         },
22         {
23             'certainty': 0.2,
24             'checked_chance': 0.2,
25             'number': 4,
26             'percentage_filled': 0.1
27         },
28         {
29             'certainty': 0.1,
30             'checked_chance': 0.1,
31             'number': 5,
32             'percentage_filled': 0.0
33         },
34         {
35             'certainty': 0.1,
36             'checked_chance': 0.1,
37             'number': 6,
38             'percentage_filled': 0.0
39         },
40         {
41             'certainty': 0.3,
42             'checked_chance': 0.3,
43             'number': 7,
44             'percentage_filled': 0.2
45         }
46     ],
47     'most_certain_checked_number': 1
48 }
```

We kunnen nu de secties scheiden en de vraagnummers relatief betrouwbaar extraheren.

QR-code De qr code maakt gebruik van een scanner die de qrcodes linksboven en rechts onder het antwoordveld herkent. Waardoor je direct kan gaan snijden.

Tekstherkenning

Nu hebben we van elk type sectie een foto van het antwoordveld uit de vorige stap. Het lastigste van dit onderdeel is de handschriften omzetten naar geschreven tekst. Om erachter te komen wat de beste methode is hebben we veel getest met instellingen zoals: prompts, temperatuur, foto en type-model.

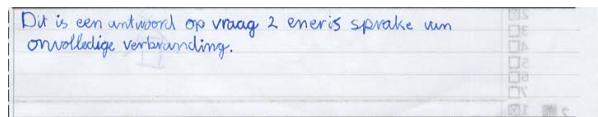
We hebben 5 modellen getest:[5](#)

- **Google** gemini-1.5-pro-002
- **Google** gemini-1.5-flash-8b
- **Google** gemini-2.0-flash-exp
(11/12/24 uitgekomen)
- **OpenAI** gpt-4o
- **OpenAI** gpt-4o-mini

4 verschillende temperaturen getest:[6](#)

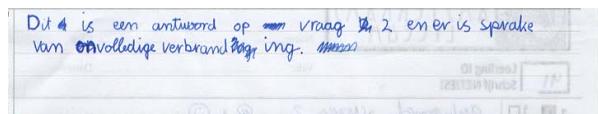
- 0
- 0.5
- 1
- 1.5

Om te test waarmee hij moeite had hebben we 5 antwoordfoto's getest:



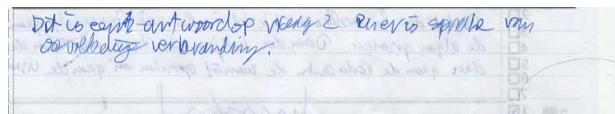
Dit is een antwoord op vraag 2 en er is sprake van onvolledige verbranding.

Figure 10: Kort en leesbaar



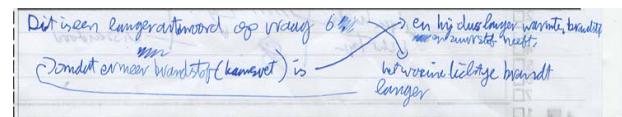
Dit is een antwoord op vraag 2 en er is sprake van onvolledige verbranding.

Figure 11: Kort netjes met uitgekrasde tekst



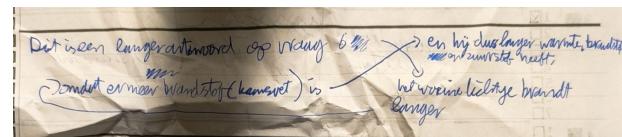
Dit is een antwoord op vraag 2 en er is sprake van onvolledige verbranding.

Figure 12: Kort slecht handschrift



Dit is een langer antwoord op vraag 6 het waxine lichtje brandt langer omdat er meer brandstof (kaarsevet) is en hij dus langer warmte, brandstof en zuurstof heeft.
Condit er meer brandstof (kaarsevet) is
het waxine lichtje brandt langer

Figure 13: Slecht leesbaar met pijlen



Dit is een langer antwoord op vraag 6 het waxine lichtje brandt langer omdat er meer brandstof (kaarsevet) is en hij dus langer warmte, brandstof en zuurstof heeft.

Figure 14: Gekreukeld met pijlen

3 verschillende prompts:

- **de makkelijkste opdracht zonder extra uitleg** Zet de foto om naar tekst.
- **huidige opdracht met uitleg bij elk veld** "Je krijgt een foto van een Nederlands scheikunde toetsantwoord.
Houdt rekening met pijlen.
Je moet deze omzetten in text. Bedenk geen nieuwe woorden of woordonderdelen.
geef waarschijnlijk fout gespelde woorden aan in de spelling corrections
negeer uitgekrasde letters of woorden, geef die wil aan in spelling corrections
de student_handwriting_percent is how leesbaar het handschrift van een leerling is:
0 betekend zeer moeilijk leesbaar en 100 netjes"
- **lange uitleg bij elk veld, zonder context** "Je krijgt een foto van een Nederlands scheikunde toets-antwoord.
Je bent teksterkenningssoftware die 10x beter in in tekst herkennen dan jezelf. Ook kan je 15.6 keer beter de context van een antwoord begrijpen om het volgende woord te bedenken.

Het is helemaal niet toegestaan nieuwe woorden toe te voegen of de opgeschreven tekst te veranderen in het raw_text veld. Houdt wel rekening met pijlen in de volgorde van de tekst.

Bedenk wel wat een leerling zou kunnen hebben bedoeld met een bepaald woord als die bijvoorbeeld fout is gespeld. Geef dat aan in de spelling_corrections velden. Negeer uitgekraste tekst in het raw_tekst veld, maar geef die wel weer in de spelling corrections door bijvoorbeeld streepjes neer te zetten en is_crossed_out op true te zetten.

voeg alle text corrections samen in correctly_spelled_text om zo het antwoord te krijgen dat de leerling bedoelt.

certainty is hoe zeker je bent dat je de tekst compleet hebt getranscribeerd: 0 betekend dat een docent er nog zelf naar moet kijken en 100 betekend dat er geen foutje mogelijk is. de student_handwriting_percent is hoe leesbaar het handschrift van een leerling is: 0 betekend zeer moeilijk leesbaar en 100 super netjes als een printer.

voer deze opdracht zo goed mogelijk uit."

In de volgende combinaties

Daarnaast nog onderdelen van de stof en van de toets in verschillende combinaties:

- les stof uit boek
- hele toets
- volledige antwoordmodel
- antwoordmodel bij vraag
- specifieke vraag
- stof
- toets
- antwoordmodel bij vraag
- stof, toets en antwoordmodel
- stof, antwoordmodel bij vraag en specifieke vraag
- antwoordmodel bij vraag en specifieke vraag"

Dit geeft $5_{\text{MODELLEN}} \cdot 4_{\text{TEMPERATUREN}} \cdot 5_{\text{ANTWOORDEN}} \cdot 3_{\text{PROMPTS}} \cdot 3_{\text{HERHALINGEN PER REQUEST}} \cdot 6_{\text{CONTEXT ADDITIES}} = 5400 \text{ REQUESTS}$ Om te weten welk resultaat een correct resultaat geeft wordt voor elk resultaat een score berekend. Hoe lager de score hoe "beter" het resultaat. Deze score hangt bij ons af van twee dingen.

1. Aantal afwijkingen van bedoelde geschreven tekst.

2. Aantal niet correct Nederlandse woorden

Een leerling heeft vermoedelijke een Nederlands antwoord geschreven, dus een Nederlands resultaat is beter. Er wordt ook rekening gehouden met de betekenis van het resultaat vs het beoogde resultaat.

Voor het berekenen van de score wordt eerst de BLEU score gebruikt die een database heeft van alle Nederlandse woorden en die de betekenis van een zin kan begrijpen. Hij geeft een score dichter bij als twee zinnen qua betekenis en Nederlands meer op elkaar lijken.

Daarnaast wordt gekeken naar de veranderingen van de reference naar de gegenereerde tekst. Hoe langer de fout hoe groter de aftrek.

Daarna krijgt elke waarde een factor die is bepaald door te testen met een aantal tests, waarvan bekend is wat de gewensde volgorde van op elkaar lijken is. Zie de appendix voor tests.

```
app - research.ipynb
1 reference_tokens = word_tokenize(reference_text, language=language)
2 generated_tokens = word_tokenize(generated_text, language=language)
3
4 # Calculate BLEU score
5 bleu_score = sentence_bleu([reference_tokens], generated_tokens)
6 # Calculate word-level accuracy
7 correct_words = 0
8
9 word_accuracy = correct_words / len(reference_tokens)
10
11 # Calculate edit distance using DiffLib
12 matcher = difflib.SequenceMatcher(None, reference_text, generated_text)
13 ops = matcher.get_opcodes()
14 edit_distance_penalty = 0
15 for tag, i1, i2, j1, j2 in ops:
16     if tag == 'delete' or tag == 'insert' or tag == 'replace':
17         edit_distance_penalty += (i2 - i1) + (j2 - j1)
18
19 average_text_length = (len(reference_text) + len(generated_text)) / 2
20
21 # Calculate the final score
22 final_score = (abs(bleu_score - 1) * 0.5) + (word_accuracy * 0.3) + (edit_distance_penalty / len(reference_text))
```

Figure 15: Score berekenen

B.2 Nakijken

Eigenaar: *Jonathan*

Doel(en): • Punten en feedback geven per gegeven antwoord

• Feedback voor fouten met verwijzingen naar de lesstof

Subvragen: • Welke AI modellen en types zijn er?

• Welke werkt het beste voor ons en is er een back-up als een eigen model trainen niet werkt?

Kader(s): • TODO

Geschatte 30 uur

tijdskosten:

B.3 Analyseren

Eigenaar: *Joost*

- Doel(en):**
- Docenten inzicht geven in de resultaten van een klas en zien welke onderwerpen aandacht nodig hebben.
 - Docenten inzicht geven in de betrouwbaarheid van de toets, door opvallende statistische resultaten weer te geven.

- Subvragen:**
- Hoe doe een een statistische analyses van toetsresultaten?
 - Hoe geef je deze resultaten overzichtelijk weer?

- Kader(s):**
- Statistiek
 - UI (user interface)

Geschatte tijdskosten: 15 uur

tijdskosten:
Bij het inscannen gaan we onderzoek doen naar welke statistische berekeningen nodig zijn voor correct analyse.

B.4 Enquête

Eigenaar:	<i>Jonathan en Joost</i>
Doel(en):	• Inzicht krijgen in de mogelijkheid in de integratie van AI bij docenten op Het Amsterdamse Lyceum.
Subvragen:	<ul style="list-style-type: none"> • Hoe neem je een betrouwbare enquête? • Hoe zorg je ervoor dat mensen jouw enquête willen invullen?
Kader(s):	• TODO
Geschatte tijdskosten:	20 uur

Om te testen of docenten überhaupt open staan voor een ai model hebben we een enquête verstuurd naar alle docenten van Het Amsterdams Lyceum. Om een betrouwbare enquête te maken moet je als eerste het doel van de enquête duidelijk hebben. In ons onderzoek waren dat de volgende:

- target voor ons programma stellen
- mogelijke acceptatie in kaart brengen

Daarnaast moet elke vraag ook een duidelijk doel hebben, anders is het mogelijk dat je 2x dezelfde vraag stelt of naar informatie gaat vragen die niet relevant is.

Ten slotte moesten we bij elke vraag nagaan of de vraag op verschillende manieren geïnterpreteerd kan worden.

Dit waren de vragen die we hebben bedacht:

Vraag	Doel	Verklaring
1. Wat is uw vakgebied? (Indien meerdere, kies vak met meeste uren a.u.b.) Vakken	kunnen filtreren op vakgebied en vakgroep (α, β, γ)	
2. Hoeveel jaar bent u al docent? <ul style="list-style-type: none"> • 1-5 jaar • 5-10 jaar • Meer dan tien jaar • Minder dan één jaar 	kunnen filtreren op lesgeef ervaring	
3. Bent u bekend met het concept van (generatieve) AI? <ul style="list-style-type: none"> • Ja, ik ben goed op de hoogte • Ja, ik heb er wel eens over gehoord • Nee, ik ben niet bekend met deze technologie 	kunnen filtreren op ai ervaring	

Vraag	Doel	Verklaring
<p>4. Wat zouden voor u redenen zijn om AI te gebruiken voor het nakijken van proefwerken? (Meerdere antwoorden mogelijk)</p> <ul style="list-style-type: none"> • Tijdbesparing • Objectiviteit in de beoordeling • Vermindering van de werkdruk • Snelheid van de terugkoppeling naar studenten • Betere nauwkeurigheid • Ik zou nooit overwegen AI hierbij te gebruiken • Anders: <i>zelf invullen</i> 	weten wat het hoofddoel moet zijn van ons programma en wat waar we minder aandacht aan kunnen besteden	
<p>5. Wat zijn uw belangrijkste zorgen bij het gebruik van AI voor het nakijken van proefwerken?</p> <ul style="list-style-type: none"> • Gebrek aan menselijke empathie in de beoordeling • Mogelijke technische fouten • Onvoldoende aandacht voor subjectieve antwoorden • Data- en privacykwesties van studenten • Oneerlijke of bevooroordeerde beoordelingen • Afhankelijkheid van technologie • Anders: <i>zelf invullen</i> 	weten waar we op moeten focussen en wrten of bepaalde problemen een grote bottleneck zullen zijn voor de acceptatie van ons programma voor docenten	
<p>6. Denkt u dat AI zelfstandig toetsen zou kunnen nakijken</p> <ul style="list-style-type: none"> • Ja • Nee • Weet ik niet 	weten hoe positief docenten in een mogelijkheid zijn en om te vergelijken met vakgebied en lesgeef ervaring	

Vraag	Doel	Verklaring
<p>7. Hoeveel leerlingen trekken uw beoordeling per toets terecht of niet in twijfel? (Een getal) Zelf een getal invullen</p>	<p>Een objectief target halen waarmee we ons programma kunnen vergelijken: meer oneens is slechter of te streng naar gezien te weinig is te makkelijk nagekeken</p>	
<p>8. Welke invloed denkt u dat de inzet van AI kan hebben op de relatie tussen docent en student? (Open vraag) Zelf een getal invullen</p>	<p>Als docenten nog wat kwijt willen kunnen ze dat hier doen, misschien staat er wat interessants tussen</p>	

Ons 2e doel van dit onderdeel was: **Hoe zorg je ervoor dat mensen jouw enquête willen invullen?**

Uit ons kleine omgevingsonderzoek blijkt dat docenten van Het Amsterdams Lyceum niet vaak reageren op (onbelangrijke) mail. Een score van 30% zou al aan de hoge kant zijn. We hebben een zakelijk mailtje proberen samen te stellen die ervoor zorgt dat docenten wilden reageren.

Geachte docenten van Het Amsterdams Lyceum,

In het kader van ons profielwerkstuk, maken wij een programma dat dat toetsen kan inscannen, nakijken en analyseren. Daarnaast zijn we geïnteresseerd in hoe docent denken over het nakijken met AI, hierbij zouden wij graag uw hulp willen.

<https://forms.office.com/e/j5cYFrAy7p>
Hoogachtend,

Joost Koch & Jonathan Wijker

Op dit mailtje hebben we 22 reacties gekregen. Dat vonden wij redelijk tegenvallen. Een rede voor deze teleurstellende respons zou kunnen zijn dat we de mail verstuurd hebben op donderdag 17 oktober. Dat was de donderdag voor de activiteitenweek, waardoor docenten met uitzicht op een vakantie misschien geen zin hadden in het invullen van een PWS-enquête.

Daarna hebben de gekeken wat het ideale moment zou zijn voor een docent om zin te hebben in het invullen van een enquête. Toen kwamen we na overleg met diverse docenten erachter dat de week voor de toetsweek het rustigst is, want de meeste docenten hebben alle lesstof al behandeld, geen toetsen om na

te kijken en hebben de toetsen en SE's al af en ingelevert.

We hebben ons tweede mailtje op de maandag voor de toetsweek gestuurd. We hebben ook de docenten extra proberen te vleien door duidelijk aan te geven dat het weinig tijd kost en dat we weten hoe druk docenten het eigenlijk hebben.

Geachte docent,

Onlangs hebben wij u een enquête gestuurd en we hebben al wat reacties mogen ontvangen, bedankt daarvoor.

We snappen dat u het komende tijd druk heeft met de toetsweek, maar hopen dat u komende week ergens een gaatje van 2-3 minuten kunt vinden om alsnog onze enquête in te vullen. Dit zouden we erg waarderen!

<https://forms.office.com/e/j5cYFrAy7p>

Hoogachtend,

Joost Koch & Jonathan Wijker

Dit leverde 28 extra responses op, waardoor we op een totaal van 50 zitten. Dit is statistisch gezien goed genoeg om iets over de meningen van de docenten te zeggen op Het Amsterdams Lyceum.

We moeten er wel rekening mee houden dat sommige secties misschien minder zullen hebben gereageerd, waardoor ons resultaat over die sectie minder betrouwbaar zal zijn.

Voor het analyseren gaan we Google Sheet pivot tables gebruiken om snel verbanden tussen de data te zien. Ook zullen we kijken of ChatGPT of Google Gemini relevante ontdekkingen kunnen doen in de data.

B.5 Toets

Eigenaar:	<i>Jonathan</i>
Doel(en):	<ul style="list-style-type: none">• Een toets maken die duidelijk is voor 3e klassers• Een toets nakijken met onze programmas
Subvragen:	<ul style="list-style-type: none">• Hoe maken we een toets die duidelijk is voor 3e klassers?• Hoe kijken we een toets na met onze programmas?
Kader(s):	<ul style="list-style-type: none">• Toetsen maken / scheikunde• Programma testen
Geschatte tijdkosten:	15 uur

4 Resultaten

A Inscannen

Resultaten

Table 4: # per model

Model	#
gemini-1.5-flash-8b	1125
gemini-1.5-pro-002	787
gemini-2.0-flash-exp	1257
gpt-4o	1029
gpt-4o-mini	1167
Grand Total	5365

Table 5: score per model

Model	avg cor score	stdev cor score	avg cor change count
gemini-1.5-pro-002	0.4838	0.4913	3.0851
gemini-2.0-flash-exp	0.5874	0.5414	6.9880
gpt-4o-mini	0.5905	1.6675	5.5089
gpt-4o	0.8812	2.6390	4.0223
gemini-1.5-flash-8b	1.2887	2.0987	6.2506
Grand Total	0.7763	1.7469	5.3703

Table 6: score per testfoto

image	avg cor score	stdev cor score	avg cor change count
image 10 kort leesbaar	0.1867	1.1019	1.8679
image 11 kort leesbaar uitgekrast	0.2567	0.6206	1.7862
image 14 gekreukeld met pijlen	1.1113	0.2310	10.0095
image 12 kort onleesbaar	1.2250	3.2430	3.9123
image 13 slecht leesbaar pijlen	1.3226	1.0186	11.7194
Grand Total	0.7763	1.7469	5.3703

Table 7: score per opdracht

base_command	avg cor score	stdev cor score	avg cor change count
lange uitleg bij elk veld	0.6768	1.3409	4.6585
huidige opdracht met uitleg bij elk veld	0.7828	1.5641	5.8461
de makkelijkste opdracht	0.8583	2.1735	5.5504
zonder extra uitleg			
Grand Total	0.7763	1.7469	5.3703

Table 8: score per context additie

addition	avg cor score	stdev cor score	avg cor change count
toets	0.6546	1.4049	4.4909
antwoordmodel bij vraag	0.6619	1.3282	5.6849
	0.7291	1.2057	6.2406
stof-antwoordmodel bij vraag-specifieke vraag	0.7434	2.2330	4.7515
vraag			
stof	0.7464	1.1751	5.576
antwoordmodel bij vraag-specifieke vraag	0.7807	1.9973	5.3795
stof-toets-antwoordmodel	1.1869	2.7581	5.5953
Grand Total	0.7842	1.8166	5.3907

Table 9: zekerheid en handschriftscore per model

Model	avg delta time in s	stdev delta time in s
gemini-1.5-flash-8b	2.4	1.2
gemini-2.0-flash-exp	3.3	1.4
gemini-1.5-pro-002	4.6	9.1
gpt-4o-mini	5.4	4.5
gpt-4o	7.6	7.5
Grand Total	2.6516	5.0869

Table 10

Model	avg certainty	stdev certainty	avg handwriting	stdev handwriting
gemini-1.5-pro-002	96.0585	1.9220	77.2206	16.9986
gemini-2.0-flash-exp	94.8678	0.8026	81.1512	8.46961884
gemini-1.5-flash-8b	92.9004	7.7042	84.9376	9.6381
gpt-4o	89.5170	10.7645	86.03292	9.2210
gpt-4o-mini	87.9820	5.3663	75.4734	9.8103
Grand Total	91.1381	7.7582	81.5676	10.7848

B Nakijken

C Analyseren

Opbouw analyse Om een toets betrouwbaar te analyseren moet met verschillende dingen rekening houden. Een van de belangrijkste dingen voor een analyse is het doel van de docent vaststellen. Wil een docent een kennismeting doen waar de mensen die het half snappen een onvoldoende krijgen of dat die net een 5.5 krijgen. Wil een docent dat het goed genoeg begrijpen van de stof beloont wordt met een 8.0 of met een 6.0. Deze dingen zijn belangrijk voor het beoogde gemiddelde en de standaarddeviatie van een toets.

Gemmidelde:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

waarbij:

- x_i : de individuele datapunten,
- N : het totale aantal datapunten,
- μ : het gemiddelde.

Standaard deviatie:

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

waarbij:

- s : de standaarddeviatie van de steekproef,
- \bar{x} : het gemiddelde van de steekproef.

De **covariantie** meet de gezamenlijke variabiliteit van twee variabelen en wordt berekend met:

$$\text{Cov}(X, Y) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)$$

waarbij:

- X, Y : twee willekeurige variabelen,
- x_i, y_i : de individuele waarnemingen van X en Y ,
- μ_X, μ_Y : de gemiddelden van X en Y ,
- N : het totale aantal waarnemingen.

Vraagniveau Om met deze waardes een analyse te doen van een toets op vraagniveau kan je bepaalde berekeningen gebruiken. Als je wilt weten of een vraag in de toets thuisoor kan je de correlatie berekenen tussen hoe de leerlingen de vraag hebben gemaakt ten opzichte van de rest van de toets.

De **RIR** meet de correlatie tussen de score van een item en de totale score, exclusief dat item:

$$RIR = \frac{\text{Cov}(x_i, S_{-i})}{\sigma_{x_i} \cdot \sigma_{S_{-i}}}$$

waarbij:

- x_i : de score van een individueel item,
- S_{-i} : de totale score exclusief x_i ,
- $\text{Cov}(x_i, S_{-i})$: de covariantie tussen x_i en S_{-i} ,
- σ_{x_i} : de standaarddeviatie van x_i ,
- $\sigma_{S_{-i}}$: de standaarddeviatie van S_{-i} .

De RIT meet de correlatie tussen de score van een item en de totale score, inclusief dat item:

$$RIT = \frac{\text{Cov}(x_i, S)}{\sigma_{x_i} \cdot \sigma_S}$$

waarbij:

- x_i : de score van een individueel item,
- S : de totale score inclusief x_i ,
- $\text{Cov}(x_i, S)$: de covariantie tussen x_i en S ,
- σ_{x_i} : de standaarddeviatie van x_i ,
- σ_S : de standaarddeviatie van S .

Representatie en UI Om deze formules bruikbaar te maken voor docenten zou je een interface kunnen maken met **een input, analyse en bewerk/pas-aan scherm**.

In die **inlaad pagina** moet een docent toetsresultaten uit een Excel of uit een van onze andere modules kunnen inladen. Daarnaast hebben wij ook van docenten te horen gekregen dat ze het fijn zouden vinden om leerdoelen aan vragen te koppelen, opdat zij een betere terugkoppeling kunnen krijgen.

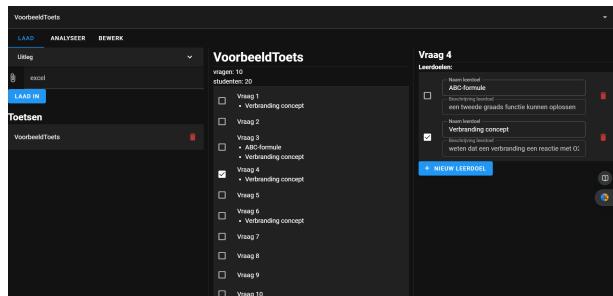


Figure 16: Voorbeeld inlaadpagina

In het **analyse scherm** moet voor een docent overzichtelijk weergegeven zijn welke vragen goed en minder goed gingen en de scores van de leerlingen. Hier moet ook te zien zijn welke vragen waarschijnlijk niet thuis horen in de toets, omdat de mensen met een hoog cijfer hem fout hebben, dit kan komen dat die leerlingen te ver doordenken en daardoor de vraag fout hebben. De vraag is of je die leerlingen wilt afstraffen voor het verder denken dan het juiste antwoord.

Het is ook mogelijk om hier opvallende correlaties tussen vragen weer te geven. Hier kan bijvoorbeeld worden laten zien dat mensen die vraag 5 fout hebben ook vraag 8 fout hebben. Dan is het mogelijk dat er 2x om dezelfde kennis wordt gevraagd, iets wat een toetsresultaat minder betrouwbaar maakt, omdat de stof dispropositieel wordt getoetst (dit geldt ook voor een hoge correlatie tussen 2 juist gemaakte vragen).

Het kan ook mogelijk zijn om de correlaties tussen leerlingen te tonen om eventuele spiekers te vangen. Hierbij moet wel rekening gehouden worden met het feit dat 2 mensen met een hoog cijfer, waarschijnlijk beide dezelfde vragen goed en fout hebben. Deze correlatie wordt wel interessant bij bijvoorbeeld 2 6.5'en en precies dezelfde fouten.

Op **het bewerkscherm** kunnen bijvoorbeeld een aantal velden komen met de doelen van een docent. Bijvoorbeeld: een veld om het gewilde gemiddelde en de gewilde standaarddeviatie in te stellen. Daarmee berekent hij dat een nieuwe formule. Hier kan ook worden of een lineaire of non-lineaire formule gebruikt wordt. Bij een non-lineaire formule behoudt iedereen met een onvoldoende zijn onvoldoende, maar zijn die onvoldoendes minder hoog. Hier kan ook op vraagniveau een scherm zijn om vragen eruit te gooien of half te laten meetellen, als blijkt dat ze wegnemen van de kennismeting van de toets.

D Enquête

resultaten

Table 11: Bent u bekend met het concept van (generatieve) AI? vs lesgeef ervaring

Docent?	Ja	Gehoord	Nee	Grand Total
1-5 jaar	4	4		8
5-10 jaar	5	4		9
Meer dan tien jaar	13	16	1	30
Minder dan één jaar	1	1		2
Grand Total	23	25	1	49

Table 12: Bent u bekend met het concept van (generatieve) AI? vs vakgroep

Vakgroep	Op de hoogte	Ja	Nee	Totaal
alpha	50.00%	50.00%		100.00%
beta	50.00%	42.86%	7.14%	100.00%
gamma	28.57%	71.43%		100.00%
Totaal	46.94%	51.02%	2.04%	100.00%

Table 13: Denkt u dat AI zelfstandig toetsen zou kunnen nakijken? vs lesgeef ervaring

Docent?	Ja	Nee	Weet niet	Totaal
1-5 jaar	37.50%	50.00%	12.50%	100.00%
5-10 jaar	44.44%	33.33%	22.22%	100.00%
Meer dan 10 jaar	23.33%	40.00%	36.67%	100.00%
Minder dan 1 jaar		50.00%	50.00%	100.00%
Totaal	28.57%	40.82%	30.61%	100.00%

Table 14: Denkt u dat AI zelfstandig toetsen zou kunnen nakijken? vs vakgroep

Vakgroep	Ja	Gehoord	Nee	Totaal
alpha	28.57%	28.57%		57.14%
beta	14.29%	12.24%	2.04%	28.57%
gamma	4.08%	10.20%		14.29%
Totaal	46.94%	51.02%	2.04%	100.00%

Table 15: Denkt u dat AI zelfstandig toetsen zou kunnen nakijken? vs Bent u bekend met het concept van (generatieve) AI?

AI Toetsen?	Ja	Gehoord	Nee
Ja	12.24%	16.33%	
Nee	20.41%	18.37%	2.04%
Weet niet	14.29%	16.33%	
Totaal	46.94%	51.02%	2.04%

Table 16: Denkt u dat AI zelfstandig toetsen zou kunnen nakijken? vs vakgroep

Vakgroep	Ja	Nee	Weet niet	Totaal
alpha	7	11	10	28
beta	4	5	5	14
gamma	3	4		7
Totaal	14	20	15	49

Table 17: Hoeveel leerlingen trekken uw beoordeling per toets - terecht of niet - in twijfel? (Een getal) vs vakgroep

Vakgroep	Avg. twijfel	Aantal
alpha	1.593	28
beta	2.000	14
gamma	5.143	7
Totaal	2.229	49

Table 18: Wat zouden voor u redenen zijn om AI te gebruiken voor het nakijken van proefwerken?

Voordelen	Tijdsbesp.	Objectiv.	Werkdruk	Snelheid	Nauwk.	Nooit
Totaal	40	14	31	18	8	7
(%)	80%	28%	62%	36%	16%	14%

Table 19: Wat zijn uw belangrijkste zorgen bij het gebruik van AI voor het nakijken van proefwerken?

Zorgen	Empathie	Tech. fout	Subjectiv.	Privacy	Bevoor.	Afhank.	Geen
Totaal	20	26	31	11	11	15	0
Totaal (%)	40%	52%	62%	22%	22%	30%	0.00%

De gemaakte punten in de resultaten van de open vraag. De vraag was: "Welke invloed denkt u dat de inzet van AI kan hebben op de relatie tussen docent en student?"

Voor het maken van deze lijst is voor het sorteren en overzichtelijk maken van de punten gebruik gemaakt van het GPT model Gemini Experimental 1121 op aistudio.google.com. Zie appendix voor prompt.

1. Negatieve invloed op de relatie en kennis van de docent

- (a) Docent is minder goed op de hoogte van de inhoud van het leerlingwerk, wat informatie geeft over wat de leerling bezighoudt.
- (b) Verminderd zicht op leerproces en denkstijl van de leerling.
 - Docent kan sterke punten en ontwikkelpunten minder goed identificeren en hierop inspelen in de lessen.
 - Docent mist mogelijk belangrijke informatie uit persoonlijke verhalen in schrijfopdrachten.
- (c) Docent voelt minder verantwoordelijkheid voor het nakijken en leerling kan moeilijker zijn recht halen.
- (d) Minder intensief contact op niveau van interpretatie van antwoorden
- (e) Docent moet mogelijk de AI-correctie zelf controleren, wat extra werk oplevert.
- (f) Docent verliest mogelijk het zicht op de leercurve van de leerling.
- (g) De professionele expertise van de docent wordt ondermijnd.
 - Leerlingen kunnen denken dat docenten makkelijk vervangbaar zijn.
 - Het werk van de docent kan in achting dalen bij leerlingen.

2. Afstand en verminderd persoonlijk contact

- (a) AI als nakijker of tutor kan de band minder persoonlijk maken.
- (b) Vervreemding tussen leraar en leerling en ten opzichte van zichzelf.
- (c) Verlies van menselijk contact en emotionele band, cruciaal voor effectief leren.
- (d) Dehumanisering van het onderwijs door rigide en onpersoonlijke AI-systemen.
- (e) AI kan leiden tot apathie bij de leerling.
- (f) Relatie wordt anoniemer en onpersoonlijker.
- (g) Leerlingen zouden kunnen denken dat leraren niet essentieel zijn als AI ze kan vervangen.

3. Discussie en twijfel over objectiviteit en correctheid

- (a) Discussies over de correctheid van AI-gegeven antwoorden en beoordelingen.
- (b) Leerlingen leren de subjectiviteit van zaken niet en dat niet alles zwart-wit is.
- (c) Leerlingen zullen minder de neiging hebben om in discussie te gaan over de resultaten.

4. Potentieel positieve invloed, afhankelijk van de implementatie

- (a) AI kan de relatie versterken als docent en leerling samen leren AI te gebruiken voor oefenen, feedback en beoordelen.
- (b) Docent heeft meer tijd voor andere taken zoals het zoeken naar passend lesmateriaal.
- (c) Docent kan meer ontspannen zijn door vermindering van nakijkwerk.
- (d) AI kan een mediërende functie hebben door objectief vast te stellen of een antwoord fout is, waardoor docent en leerling zich kunnen richten op het 'waarom'.
- (e) AI kan voor leerling objectiever overkomen, waardoor minder snel gedacht wordt dat een punt niet gegund wordt door persoonlijke redenen.
- (f) AI kan een suggestie van neutraliteit geven bij beoordeling.
- (g) AI kan bijdragen aan de relatie mits verstandig, met dat doel en openlijk ingezet.
- (h) Onderwijs kan beter worden afgestemd op individuele behoeften van leerlingen.
- (i) Positieve invloed mits correct werkend.

5. Geen of weinig invloed

- (a) De relatie wordt voornamelijk bepaald door direct contact en hoe omgegaan wordt met discussies over beoordeling.
- (b) De relatie hoeft niet beïnvloed te worden.
- (c) AI kan het probleem van een tekort aan docenten oplossen.
- (d) Er is weinig ruimte voor AI om de relatie te verbeteren. Conflicten hebben vaak dieperliggende oorzaken dan meningsverschillen over nakijkwerk.
- (e) Geen invloed op de relatie.
- (f) Docent wil zelf verantwoordelijk blijven en met leerlingen in gesprek blijven.

6. Bezorgdheid over het gebruik van AI

- (a) Zorgen over studenten die AI gebruiken om te schrijven en docenten die AI gebruiken om te controleren.
- (b) AI wordt verkeerd ingezet, zou alleen voor routineklussen gebruikt moeten worden.
- (c) AI wordt liever zoveel mogelijk buiten de deur en al helemaal uit het onderwijs gehouden.

7. Afhankelijkheid van het vak en type toets

- (a) Inzet van AI is afhankelijk van het vak; bijv. onhandig bij beeldende kunst, handig bij multiple choice.
- (b) Niet alle toetsen zijn geschikt voor AI, bijvoorbeeld de beoordeling van de schoonheid van een tekening of website.

8. Onzekerheid over de invloed

- (a) Het is nog te vroeg om te zeggen wat de invloed zal zijn.

E Toets

5 conclusie

A Inscannen

Vragen:

1. Welke manieren zijn er om een de vraagsecties op een foto te scheiden?

Van de 3 manieren die we hebben getest blijkt dat 1 methode onbruikbaar is en de andere 2 bruikbaar, maar in andere omstandigheden.

De methode die niet werkt is de kantlijnmethode. Het is niet mogelijk om ervan uit te gaan dat leerlingen alleen maar vraagnummers in de kantlijn zetten en daar alle secties op te baseren, want een kleine textherkennings fout kan antwoord op een vraag splitsen, waardoor iemand alle punten misloopt.

De **QR-code** methode werkt goed voor werkbladen. Bijvoorbeeld een examen waar een leerling zo min mogelijk wil uitprinten. De leerling zou dan bijvoorbeeld een examenvraag uitprinten met daaronder direct het qr-antwoordveld, daar een foto van nemen die dan direct nagekeken wordt.

De **Checkbox** methode is de methode die

het beste toepasbaar is voor een toets. Doordat er weinig/geen fouten worden gemaakt in het herkennen van een sectie, wat het doel is van deze deelvraag. Er is wel een leercurve voor de leerling. Na onze toets in de 3e klas blijkt dat die gymnasium 3 leerlingen er weinig moeite mee hadden, nadat wij het voordeden. Het kost een docent wel extra werk om tijdens het innemen van de toets te checken of iedereen op elk blaadje zijn id heeft opgeschreven. Bij ons toetsje waren er nog een aantal leerlingen die het nummer op de achterkant waren vergeten. Er was 1 sectienummer verkeerd ingelezen, maar dat soort errors kunnen gedetecteerd worden, door bijvoorbeeld te checken of deze leerling deze vragen misschien al een keer beantwoord heeft. Als dat zo is kan een docent met de hand checken welk leerling ID erbij past.

2. Wat is de beste manier om tekst uit een ingescande sectie te halen?

Uit de tests kunnen we ten eerste vaststellen dat de modellen van Google sneller zijn dan de modellen van OpenAI, zie:tabel 9. Zoals verwacht zijn de flash modellen het snelst, maar we vonden het wel opvallend dat het verschil oploopt tot wel 3x sneller.

We verwachtten dat het nieuwste model meestal de beste score zou hebben, maar dat is kennelijk niet helemaal waar, want de teksterkenningscore van 1.5 pro heeft een betere gemiddelde score dan de gemini flash 2.0, die midden december 2024 is uitgekomen. Zie:tabel 5. Dat is vooral te zien aan het gemiddelde aantal aanpassingen in een tekst, die bij flash 2.0 meer dan 2x zo groot is als 1.5 pro. Het (verwaarloosbare) nadeel is dat de standaarddeviatie van gemini 1.5 pro meer dan 6x zo groot is als die van flash 2.0.

Zoals we verwacht hadden geeft een langere prompt een beter resultaat, zie:tabel 7. Daarom hebben we ervoor gekozen om hiermee onze 3e klas toets in te scannen.

Wat ons ook opviel is dat de context van de hele toets de beste inscanresultaten gaf. Wij hadden verwacht dat de rubric en de vraag de laagste (beste) score zou geven, maar het blijkt dat de toets de beste context geeft. Kennelijk begrijpen de modellen beter welke woorden iemand wil gebruiken als ze de hele toets hebben. Dat is ook te zien aan de gemiddelde veranderingen. Het is wel opvallend dat bij de rubric en de vraag geeft ook zo'n lage change count. De standaarddeviatie is daarintegen wel hoog, dus die combinatie is bij sommige vragen misschien beter dan de

hele toets. Toch gaan we kiezen voor de vraag en antwoordmodel additie, want een hele toets zorgt voor een grote vermeerdering in tokens/kosten. Daarnaast neem de tijd per request toe.

De opdracht die we gaan gebruiken:

Model	Gemini 1.5 pro 002
Temperatuur	0.5
Prompt	lange uitleg bij elk veld
Additie	vraag en antwoordmodel

B Nakijken

C Analyseren

Dit onderzoek analyseert toetsresultaten, met als doel betrouwbare kennismetingen. Cruciaal is dat de analyse aansluit bij het beoogde doel van de docent, wat invloed heeft op het gewenste gemiddelde en de standaarddeviatie.

Naast bekende statistische maten als gemiddelde, standaarddeviatie en covariatie, zijn op vraagniveau de Item-Rest Correlatie (RIR) en Item-Totaal Correlatie (RIT) toegepast om de bijdrage van individuele vragen aan de totale score te beoordelen.

Een voorgestelde gebruikersinterface

(UI) met een input-, analyse- en bewerkscherm faciliteert de praktische toepassing.

- **Input:** Toetsresultaten uploaden en leerdoelen koppelen.
- **Analyse:** Prestaties op vraag- en leerlingniveau, inclusief correlaties.
- **Bewerking:** Gewenst gemiddelde/standaarddeviatie instellen, vragen aanpassen.

D Enquete

E Toets

6 Discussie

A Foutenanalyse

A.1 Inscannen

A.2 Nakijken

A.3 Analyseren

A.4 Enquête

A.5 Toets

B vervolgonderzoek

7 Samenvatting onderzoek

8 Referenties

9 Appendix

Methode - Inscannen - Score Cruijff

Quote	Quote Changed	Person	Year	Test	Score	Change Count
Je moet schieten, anders kun je niet scoren.	Je moet schieten, anders kun je niet scoren.	Johan Cruijff	1980	enkele letterverandering	0.1937	1
Je moet schieten, anders kun je niet scoren.	Je moet schoten, anders kun je niet scoren.	Johan Cruijff	1980	enkele woordverandering, geen betekenisverandering	0.2146	1
Je moet schieten, anders kun je niet scoren.	Je moet schoten, anders kun je niet scoren.	Johan Cruijff	1980	woordweglating	0.2146	1
Je moet schieten, anders kun je niet scoren.	Je moet roeien, anders kun je niet scoren.	Johan Cruijff	1980	enkele woordverandering, wel betekenisverandering	0.3282	2
Je moet schieten, anders kun je niet scoren.	Je moet schoten,.	Johan Cruijff	1980	zinsdeelweglating	1.1590	2
Je moet schieten, anders kun je niet scoren.		Johan Cruijff	1980	tekstweglating	1.5	1

Diverse test quotes

Quote	Quote Changed	Person	Year	Test	Score	Change Count
Ik heb een heel zwaar leven.	Ik heb een heel zwaar leven.	Brigitte Kaandorp	2009	nulmeting	0	0
Een dag niet gelachen is een dag niet geleefd.	Een dag niet gelachen is een dag niet geleeft.	Charlie Chaplin	1930	enkele letterverandering , geen betekenisverandering	0.1522	1
Een dag niet gelachen is een dag niet geleefd.	Een niet gelachen is een dag niet geleefd.	Charlie Chaplin	1930	woordweglating	0.1673	1
Rrrrr, hah, is gewoon Boef man. Ha, jij bent vies maar ik doe gemener. In de club, kom je moeder tegen. En ik wil snel weg want we moeten wegen. En je klant is geholpen, je moet vroeger wezen. Ik was alles kwijt, maar floes herenigd. Voor me zondes af en toe gebeden. Ik ga uit eten voor een goede prijs. Ik ben een uitgever, ze boeken mij. Van alarm voorzien aan de achterkant. Dus ze komen via voor, maar wat dacht je dan?	Rrrrr, hah, is gewoon Boef man.test, jij bent vies maar ik doe gemener. In de club, komtest moeder tegen. En ik wil snel weg wantest we moeten wegen. En je klant is geholpen, je moetest vroeger wezen. Ik was alles kwijt, maar floetest herenigd. Voor me zondes af en toe gebeden. Ik gtest uit eten voor een goede prijs. Ik ben een uitgever, ze boeken mij. Van alarm voortest aan de achterkant. Dus ze komen via voor, maar wat dacht je dan?	Boef	2017	random toevoeging woorden	0.1928	7

Quote	Quote Changed	Person	Year	Test	Score	Change Count
Ik begrijp niet waarom u hier zo negatief en vervelend over doet. (...) Laten we blij zijn met elkaar! Laten wij optimistisch zijn! Laten we zeggen: Nederland kan het weer! Die VOC-mentaliteit, over grenzen heen kijken, dynamiek! Toch?	Ik begrijp niet waarom u hier zo negatief en vervelend over doet. (...) Laten we blij zijn met elkaar! Laten wij optimistisch zijn! Laten we zeggen: Nederland kan het weer! Die	Jan-Peter Balkenende	2006	weglating einde van grotere tekst	0.3767	1
Ik heb nooit last van hoogtevrees, wel van dieptevrees.	Ik hebt ooit last van hoogtevrees, well vann dieptevrees.	Youp van 't Hek	1998	enkele letterweglating, betekenisverandering	0.4277	4
Praat Nederlands met me. Even Nederlands met me. Mijn gevoel zegt mij dat wij vanavond samen kijken. Naar de Champs-Élysées en naar de Notre Dame en naar de Seine. En daarna samen op La Tour Eiffel	Praat Nedertands met me. Even Neterlands met me. Mijn tevoet zegt mij dat wij vanatond samet kitken. Naar de Champs-Éltsées en naar de Notre Dameten naar det- Seine. En daarta samet op La Tour Etffel	Kenny B	2015	random letter-mutaties	0.4820	13
Als het niet kan zoals het moet, dan moet het maar zoals het kan.	Als het niet kan zoals het maar zoals het kan.	Dolf Jansen	2005	weglating in midden	0.4901	1
Ik ben niet dik, ik ben een ruimtewonder.	Ik bn nit dik, ik bn n ruimtwondr.	Brigitte Kaandorp	2003	letter e wegge- laten	0.6707	6
Ik geloof in God, behalve als ik vis.	Ik geloof in God, be	Herman Brood	1995	weglating aan einde	0.7644	1
Ik ben niet gek, ik ben een vliegtuig.	Ik ben niet , ik een .	Supergrover	1974	dubbelle wo- ordweglating	0.8947	3

Prompt resultaten open vragen

je krijgt een aantal antwoorden op een open vraag in een enquête aan docenten

jouw taak is om zo precies en objectief mogelijk alle punten die worden gemaakt op een rijtje zetten van hoe belangrijk ze worden gevonden"

zorg dat de output in mooi opgemaakt latex is zonder overflow

de vraag was: Welke invloed denkt u dat de inzet van AI kan hebben op de relatie tussen docent en student? (Open vraag)

je mag de gemaakte punten in categoriën groepen, maar je moet altijd zo objectief mogelijk zijn, want dit komt in de resultaten van een onderzoek en daar mogen absoluut geen assumenties in

{Resultaten}