|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Iteratie** | **Omschrijving** | **Prestaties** | | **Wat opvalt** |
| 1 | Combinatie objectdetectie en HAR toegevoegd aan de transcriptie en captions die al bestonden | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.600000 | 0.0 | 0.220892 | | 0.666667 | 0.0 | 0.224851 | | 1.000000 | 0.0 | 0.218888 | | 0.750000 | 0.0 | 0.220406 | | 0.500000 | 0.0 | 0.221987 | | 0.666667 | 0.0 | 0.224239 | | 0.500000 | 0.0 | 0.356924 | | 0.000000 | 0.0 | 0.192590 | | 0.000000 | 0.0 | 0.227117 | | 0.266667 | 0.0 | 0.220827 |  * Gemiddelde faithfulness: 0.495 * Gemiddelde answer\_relevancy: 0.0 * Gemiddelde answer\_correctness: 0.232872 | | Bij het opslaan van de valuable knowledge valt op dat er weinig informatie wordt opgeslagen en wat erop wordt geslagen komt bijna uitsluitend van de transcriptie. |
| 2 | Prompt van de knowledge extractor aangepast. Om meer informatie op te slaan.  Aangepast van: “You are a knowledge extractor. Return ONLY JSON with the following schema:”  Naar:  "You are a knowledge extractor. Extract subject-predicate-object knowledge triplets from all the provided information: transcript, caption and linked object-action pairs. Return ONLY JSON with the following schema:" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.000000 | 0.220299 | | 0.833333 | 0.000000 | 0.220727 | | 0.750000 | 0.000000 | 0.204240 | | 0.666667 | 0.915188 | 0.721596 | | 0.285714 | 0.000000 | 0.223569 | | 0.875000 | 0.000000 | 0.216304 | | 0.000000 | 0.000000 | 0.209285 | | 0.400000 | 0.000000 | 0.564960 | | 0.500000 | 0.887634 | 0.442737 | | 0.333333 | 0.000000 | 0.215304 |  * Gemiddelde faithfulness: 0.564405 * Gemiddelde answer\_relevancy: 0.180282 * Gemiddelde answer\_correctness: 0.323902 | | Het HAR model detecteert weinig acties waardoor er ook weinig objecten worden opgeslagen in de knowledge graph die wel relevant zijn. |
| 3 | Handdetectie model toegevoegd. Veel relevante objecten werden niet toegevoegd aan de knowlegde graph omdat de combinatie objectdetetcie en HAR niet goedwerkte vanwege de prestaties van het HAR model. De input van het framework zullen technische video’s zijn waar altijd met de handen gewerkt wordt. Objecten die worden aangeraakt zullen waarschijnlijk relevant zijn in de video en voor de vragen die erover gesteld worden. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.571429 | 0.000000 | 0.219198 | | 1.000000 | 0.000000 | 0.230695 | | 1.000000 | 0.000000 | 0.210944 | | 0.833333 | 0.000000 | 0.414248 | | 0.800000 | 0.000000 | 0.339856 | | 0.571429 | 0.000000 | 0.722578 | | 0.400000 | 0.903935 | 0.582781 | | 0.000000 | 0.962928 | 0.190841 | | 1.000000 | 0.978993 | 0.239745 | | 0.000000 | 0.000000 | 0.214675 |  * Gemiddelde faithfulness: 0.617619 * Gemiddelde answer\_relevancy: 0.284585 * Gemiddelde answer\_correctness: 0.336556 | | Bij het opslaan van de informatie in de knowledge graph valt op de combinatie van de objectdetectie en HAR niet altijd goed gaat.  {'subject': 'Kabel', 'predicate': 'is holding', 'object': 'something'} |
| 4 | De objecten invullen op de plek van de somthing van HAR. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.857143 | 0.000000 | 0.225605 | | 0.600000 | 0.000000 | 0.224139 | | 0.333333 | 0.000000 | 0.215768 | | 0.545455 | 0.000000 | 0.219744 | | 0.166667 | 0.941671 | 0.222317 | | 0.250000 | 0.936857 | 0.470008 | | 0.200000 | 0.000000 | 0.340553 | | 0.000000 | 0.948460 | 0.186688 | | 0.636364 | 0.000000 | 0.220582 | | 0.250000 | 0.000000 | 0.222059 |  * Gemiddelde faithfulness: 0.383896 * Gemiddelde answer\_relevancy: 0.282699 * Gemiddelde answer\_correctness: 0.254746 | | Er wordt nu minder opgeslagen in de knowledge graph. Ik denk dat deze iteratie potentie heeft maar dat eerst de promp aangepast moet worden  LLM een keer Chinees. |
| 5  Vervolg op iteratie 3 | Touche weghalen achter de objecten die aangeraakt worden. Die woorden namelijk nu zo opgeslagen: {'subject': 'een hand', 'predicate': 'toucheert', 'object': 'een Kabel'} | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.750000 | 0.000000 | 0.219467 | | 1.000000 | 0.000000 | 0.223479 | | 0.500000 | 0.000000 | 0.463814 | | 0.333333 | 0.841337 | 0.519889 | | 0.857143 | 0.000000 | 0.215762 | | 1.000000 | 0.000000 | 0.441808 | | 0.416667 | 0.000000 | 0.199764 | | 0.400000 | 0.000000 | 0.188084 | | 1.000000 | 0.000000 | 0.225858 | | 0.428571 | 0.000000 | 0.212159 |  * Gemiddelde faithfulness: 0.668571 * Gemiddelde answer\_relevancy: 0.084134 * Gemiddelde answer\_correctness: 0.291009 | | De relevante objecten worden nu bijna niet meer toegevoegd aan de valuable knowledge. |
| 6  Vervolg op iteratie 3 | De input van het model is deel Engels en deels Nederlands. Dit kan mogelijk voor verwarring zorgen bij het model. Daarom wordt de prompt van de knowledge extraxter vertaald naar het nerderlands en wordt er gevraagd om alles naar het nederlands te vertalen voor het opslaan. Nieuwe prompt: Je bent een kennis-extractor. Extraheer subject-predicate-object knowledge triplets uit alle aangeleverde informatie: transcript, caption, linked object-action pairs and hand-object interactions.  Vertaal voor het opslaan alles naar het Nederlands.  Geef ALLEEN JSON terug volgens het volgende schema:  Prompt van het LLM dat antwoord geeft vertaald naar het NL: Je bent een behulpzame assistent;  De huidige datum en tijd is: {current\_time} Antwoord op basis van opgehaalde kennis: | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.500000 | 0.960537 | 0.228736 | | 1.000000 | 0.000000 | 0.206549 | | 0.000000 | 0.911851 | 0.212059 | | 1.000000 | 0.936677 | 0.230308 | | 0.090909 | 0.000000 | 0.344024 | | 0.428571 | 0.000000 | 0.211146 | | 0.111111 | 0.000000 | 0.206554 | | 0.250000 | 0.000000 | 0.189989 | | 0.833333 | 0.000000 | 0.228116 | | 0.565217 | 0.000000 | 0.215498 |  * Gemiddelde faithfulness: 0.477914 * Gemiddelde answer\_relevancy: 0.280906   Gemiddelde answer\_correctness: 0.227298 | | Het model presteerde slechter. |
| 7  Vervolg op iteratie 3 | Er wordt veel informatie niet opgeslagen in de knowledge graph die mogelijk wel relevant kan zijn. Daarom wordt er deze zin toegevoegd aan de prompt van de knowledge extracter “Be thorough in extracting information.  ” | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.951266 | 0.230859 | | 0.800000 | 0.000000 | 0.226043 | | 1.000000 | 0.000000 | 0.211148 | | 1.000000 | 0.000000 | 0.219055 | | 0.666667 | 0.000000 | 0.223634 | | 1.000000 | 0.000000 | 0.214332 | | 0.500000 | 0.000000 | 0.334993 | | 0.222222 | 0.000000 | 0.192345 | | 1.000000 | 0.000000 | 0.608509 | | 0.500000 | 0.000000 | 0.219770 |  * Gemiddelde faithfulness: 0.768889 * Gemiddelde answer\_relevancy: 0.095127 * Gemiddelde answer\_correctness: 0.268069 | | Er werd veel meer informatie aan de knowledege graph toegevoegd maar dit zorgde niet voor betere antwoorden van het LLM. |
| 8  Vervolg op iteratie 3 | Verschillende LLM’s proberen die mogelijk beter presteren dan het Qwen model | **Mistral-7B-Instruct-v0.1.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.935198 | 0.225705 | | 0.0 | 0.952078 | 0.227784 | | 0.0 | 0.981520 | 0.227077 | | 0.0 | 0.905360 | 0.220645 | | 0.0 | 0.947943 | 0.229956 | | 0.0 | 0.955562 | 0.229966 | | 0.0 | 0.911180 | 0.201968 | | 0.0 | 0.962928 | 0.190729 | | 0.0 | 0.945237 | 0.229687 | | 0.0 | 0.934434 | 0.217307 |  * Gemiddelde faithfulness: 0.0 * Gemiddelde answer\_relevancy: 0.943144 * Gemiddelde answer\_correctness: 0.220082   **Llama-3.2-3B-Instruct-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.865706 | 0.236070 | | 0.000000 | 0.978145 | 0.237612 | | 0.000000 | 0.000000 | 0.211212 | | 1.000000 | 0.987900 | 0.988076 | | 0.000000 | 0.000000 | 0.200497 | | 0.000000 | 0.810202 | 0.218826 | | 0.200000 | 0.890193 | 0.590791 | | 0.000000 | 0.962928 | 0.193850 | | 0.500000 | 0.961085 | 0.747013 | | 0.411765 | 0.937211 | 0.220070 |  * Gemiddelde faithfulness: 0.211176 * Gemiddelde answer\_relevancy: 0.739337 * Gemiddelde answer\_correctness: 0.384402   **openchat-3.5-0106.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.166667 | 0.907987 | 0.225559 | | 0.666667 | 0.000000 | 0.219301 | | 0.000000 | 0.938904 | 0.220652 | | 0.333333 | 0.996025 | 0.603278 | | 0.000000 | 0.956245 | 0.999210 | | 0.000000 | 0.896605 | 0.226454 | | 0.000000 | 0.871358 | 0.202532 | | 0.000000 | 0.962168 | 0.192077 | | 0.142857 | 0.919894 | 0.453735 | | 0.000000 | 0.945441 | 0.221701 |  * Gemiddelde faithfulness: 0.130952 * Gemiddelde answer\_relevancy: 0.839463 * Gemiddelde answer\_correctness: 0.35645   **deepseek-llm-7b-chat.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.500000 | 0.851819 | 0.216033 | | 0.000000 | 0.922216 | 0.226919 | | 0.000000 | 0.894831 | 0.217591 | | 0.200000 | 0.987900 | 0.446757 | | 0.000000 | 0.956245 | 0.237757 | | 0.000000 | 0.955562 | 0.240911 | | 0.000000 | 0.868911 | 0.204265 | | 0.000000 | 0.962928 | 0.192035 | | 0.250000 | 0.978993 | 0.541320 | | 0.428571 | 1.000000 | 0.216177 |  * Gemiddelde faithfulness: 0.137857 * Gemiddelde answer\_relevancy: 0.93794 * Gemiddelde answer\_correctness: 0.273977   **Yi-1.5-9B-Chat-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.333333 | 0.942058 | 0.226062 | | 0.000000 | 0.984372 | 0.437430 | | 0.000000 | 0.882667 | 0.211544 | | 0.400000 | 0.854006 | 0.438602 | | 0.000000 | 0.836760 | 0.523498 | | 0.000000 | 0.000000 | 0.194152 | | 0.000000 | 0.823568 | 0.190147 | | 0.000000 | 0.927949 | 0.189546 | | 0.166667 | 0.978993 | 0.453999 | | 0.000000 | 0.846924 | 0.194935 |  * Gemiddelde faithfulness: 0.09 * Gemiddelde answer\_relevancy: 0.80773 * Gemiddelde answer\_correctness: 0.305992   **sciphi-self-rag-mistral-7b-32k.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.000000 | 0.202799 | | 1.000000 | 0.878557 | 0.223623 | | 0.000000 | 0.874029 | 0.219021 | | 0.000000 | 0.979656 | 0.224842 | | 1.000000 | 0.000000 | 0.214317 | | 0.000000 | 0.971115 | 0.223458 | | 0.000000 | 0.722794 | 0.841917 | | 0.500000 | 0.841636 | 0.194330 | | 0.333333 | 0.863889 | 0.589837 | | 0.333333 | 0.000000 | 0.218227 |  * Gemiddelde faithfulness: 0.316667 * Gemiddelde answer\_relevancy: 0.613168 * Gemiddelde answer\_correctness: 0.315237   **Meta-Llama-3.1-8B-Instruct-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.929172 | 0.236122 | | 0.000000 | 0.948121 | 0.230020 | | 0.000000 | 0.878080 | 0.227678 | | 0.500000 | 0.000000 | 0.601620 | | 0.000000 | 0.841278 | 0.228899 | | 0.000000 | 0.000000 | 0.217750 | | 0.333333 | 0.000000 | 0.512896 | | 0.000000 | 0.828686 | 0.195019 | | 0.500000 | 0.978993 | 0.741911 | | 0.176471 | 0.000000 | 0.217033 |  * Gemiddelde faithfulness: 0.15098 * Gemiddelde answer\_relevancy: 0.540433 * Gemiddelde answer\_correctness: 0.340895   **DeepSeek-R1-Distill-Qwen-14B-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.333333 | 0.929162 | 0.227878 | | NaN | 0.000000 | 0.180507 | | 0.200000 | 0.946638 | 0.218755 | | NaN | 0.000000 | 0.181692 | | 0.357143 | 0.977849 | 0.406116 | | 0.600000 | 0.925208 | 0.227373 | | 0.117647 | 0.921662 | 0.367127 | | 0.071429 | 0.962928 | 0.191510 | | 1.000000 | 0.939462 | 0.998787 | | 0.142857 | 0.924313 | 0.376342 |  * Gemiddelde faithfulness: 0.313601 * Gemiddelde answer\_relevancy: 0.836358 * Gemiddelde answer\_correctness: 0,375121 | | Van alle model gaf Llama-3.2-3B-Instruct-Q4\_K\_M het beste, korte duidelijke antwoorden. |
| 9 | In verschillende papers wordt aangeraden om het LLM een persona mee te geven.  Daarom is deze zin toegevoegd: “Your task is to answer questions from technicians about technical videos.” | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.929568 | 0.236564 | | 0.0 | 0.959654 | 0.235737 | | 0.0 | 0.997366 | 0.229959 | | 0.5 | 0.996987 | 0.603534 | | 0.0 | 0.956245 | 0.233563 | | 1.0 | 0.761531 | 0.208429 | | 0.0 | 0.844731 | 0.216330 | | 0.0 | 0.879982 | 0.190724 | | 0.5 | 0.819458 | 0.721074 | | 0.0 | 0.926232 | 0.221567 |  * Gemiddelde faithfulness: 0.2 * Gemiddelde answer\_relevancy: 0.907175 * Gemiddelde answer\_correctness: 0.309748 | |  |
| 10  Vervolg op iteratie 8 | Het publiek toevoegen aan de prompt.  “You provide answers to technicians.” Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.929568 | 0.236564 | | 0.500000 | 0.920484 | 0.225617 | | 0.000000 | 0.997366 | 0.229959 | | 0.500000 | 1.000000 | 0.603534 | | 0.000000 | 0.956245 | 0.223419 | | 0.000000 | 0.944732 | 0.231980 | | 0.000000 | 0.922812 | 0.459823 | | 0.000000 | 0.962928 | 0.191985 | | 0.500000 | 0.961085 | 0.745870 | | 0.071429 | 0.937211 | 0.221402 |  * Gemiddelde faithfulness: 0.157143 * Gemiddelde answer\_relevancy: 0.953243 * Gemiddelde answer\_correctness: 0.337015 | | Het model geeft nu goed antwoord op de vraag, maar hallucineert meer. |
| 11  Vervolg op iteratie 8 | In de informatie waar triplets van gemaakt worden zit nog veel ruis waardoor het LLM moeite heeft met de kennis te extraheren. Daarom wordt er eerst een samenvatting gemaakt van alle informatie voor dat de kennis wordt geëxtraheerd. Hierbij worden ook weer de verschillende taalmodellen geprobeerd. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.0 | 0.926179 | 0.231462 | | 0.0 | 0.991470 | 0.241110 | | 0.0 | 0.874029 | 0.228474 | | 0.0 | 0.987900 | 0.226262 | | 0.0 | 0.944867 | 0.993779 | | 0.0 | 0.950246 | 0.230085 | | 0.0 | 0.977336 | 0.212110 | | 0.4 | 0.962928 | 0.764184 | | 0.0 | 0.968666 | 0.228351 | | 0.0 | 0.913079 | 0.217959 |   **Qwen2.5-7B-Instruct-Q4\_K\_M**   * Gemiddelde faithfulness: 0.14 * Gemiddelde answer\_relevancy: 0.94967 * Gemiddelde answer\_correctness: 0.357378   **Mistral-7B-Instruct-v0.1.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.0 | 0.768127 | 0.208004 | | 0.0 | 0.991470 | 0.241110 | | 0.0 | 1.000000 | 0.223025 | | 0.0 | 0.955004 | 0.223343 | | 0.0 | 0.000000 | 0.192586 | | 0.0 | 0.955562 | 1.000000 | | 0.5 | 0.911680 | 0.197894 | | 0.0 | 0.962928 | 0.190867 | | 0.0 | 0.966285 | 0.233855 | | 0.0 | 0.939830 | 0.217272 |  * Gemiddelde faithfulness: 0.15 * Gemiddelde answer\_relevancy: 0.845088 * Gemiddelde answer\_correctness: 0.292796   **Llama-3.2-3B-Instruct-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.000000 | 0.199205 | | 0.000000 | 0.991366 | 0.234971 | | 0.000000 | 0.958010 | 0.227334 | | 0.500000 | 0.957380 | 0.228752 | | 0.000000 | 0.956245 | 0.233556 | | 0.000000 | 0.955562 | 0.234083 | | 0.000000 | 0.867828 | 0.206323 | | 0.000000 | 0.962928 | 0.188800 | | 0.666667 | 0.000000 | 0.525466 | | 0.000000 | 0.933892 | 0.214725 |  * Gemiddelde faithfulness: 0.116667 * Gemiddelde answer\_relevancy: 0.758321 * Gemiddelde answer\_correctness: 0.249322   **openchat-3.5-0106.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.500000 | 0.977864 | 0.609386 | | 0.000000 | 0.991470 | 0.237995 | | 0.000000 | 0.905973 | 0.221047 | | 0.800000 | 0.000000 | 0.217584 | | 0.000000 | 0.830110 | 0.712304 | | 0.000000 | 0.000000 | 0.222670 | | 0.000000 | 0.972413 | 0.719156 | | 0.000000 | 0.910588 | 0.495622 | | 0.000000 | 0.842622 | 0.220617 | | 0.117647 | 0.933892 | 0.216378 |  * Gemiddelde faithfulness: 0.141765 * Gemiddelde answer\_relevancy: 0.736493 * Gemiddelde answer\_correctness: 0.387276   **deepseek-llm-7b-chat.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.00 | 0.881400 | 0.228101 | | 0.00 | 0.875675 | 0.220732 | | 0.00 | 0.858568 | 0.230872 | | 0.00 | 0.000000 | 0.214001 | | 0.00 | 0.786542 | 0.204397 | | 0.00 | 0.876941 | 0.236436 | | 0.75 | 0.932013 | 0.217998 | | 0.25 | 0.828316 | 0.198029 | | 0.00 | 0.952392 | 0.237466 | | 0.00 | 0.934434 | 0.217264 |  * Gemiddelde faithfulness: 0.1 * Gemiddelde answer\_relevancy: 0.792628 * Gemiddelde answer\_correctness: 0.22053   **Yi-1.5-9B-Chat-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.970574 | 0.230796 | | 0.0 | 0.917877 | 0.234774 | | 0.0 | 0.866300 | 0.222803 | | 0.0 | 0.858188 | 0.225834 | | 0.0 | 0.956245 | 0.233556 | | 0.0 | 0.000000 | 0.220849 | | 1.0 | 0.841227 | 0.597769 | | 0.0 | 0.828316 | 0.196878 | | 0.0 | 0.968666 | 0.235671 | | 0.0 | 0.952006 | 0.217902 |  * Gemiddelde faithfulness: 0.1 * Gemiddelde answer\_relevancy: 0.81594 * Gemiddelde answer\_correctness: 0.261683   **sciphi-self-rag-mistral-7b-32k.Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.873098 | 0.219503 | | 0.0 | 0.991470 | 0.236142 | | 0.0 | 0.000000 | 0.208383 | | 1.0 | 0.842743 | 0.224625 | | 0.0 | 0.000000 | 0.222979 | | 0.0 | 0.000000 | 0.184622 | | 0.0 | 0.000000 | 0.196558 | | 0.0 | 0.962928 | 0.196548 | | 0.5 | 0.781169 | 0.221139 | | 0.0 | 0.937506 | 0.215933 |  * Gemiddelde faithfulness: 0.15 * Gemiddelde answer\_relevancy: 0.538892 * Gemiddelde answer\_correctness: 0.212643   **Meta-Llama-3.1-8B-Instruct-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.995533 | 0.230069 | | 0.000000 | 0.971482 | 0.241455 | | 0.333333 | 0.934386 | 0.222337 | | 0.500000 | 0.987900 | 0.228379 | | 0.000000 | 0.000000 | 0.194349 | | 1.000000 | 0.000000 | 0.211561 | | 0.250000 | 0.000000 | 0.194624 | | 0.000000 | 0.976137 | 0.189857 | | 0.500000 | 0.897715 | 0.611099 | | 0.000000 | 0.943549 | 0.219338 |  * Gemiddelde faithfulness: 0.258333 * Gemiddelde answer\_relevancy: 0.67067 * Gemiddelde answer\_correctness: 0.254307   **DeepSeek-R1-Distill-Qwen-14B-Q4\_K\_M**   |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.333333 | 0.000000 | 0.219145 | | 0.363636 | 0.000000 | 0.223313 | | 1.000000 | 0.876155 | 0.220120 | | 1.000000 | 0.987900 | 0.227862 | | 0.000000 | 0.792081 | 0.966554 | | 0.000000 | 0.955562 | 0.235262 | | 0.000000 | 0.917529 | 0.515831 | | 0.333333 | 0.962928 | 0.194284 | | 0.000000 | 0.783437 | 0.216421 | | 0.000000 | 0.932894 | 0.212953 |  * Gemiddelde faithfulness: 0.30303 * Gemiddelde answer\_relevancy: 0.720849 * Gemiddelde answer\_correctness: 0.323174 | | Het openchat model scoort iets beter dan het model uit iteraties 8. Echter zagen de samenvattingen er niet goed uit en lijkt het meer op toeval dat het model nu goed presteerde. Daarom wordt er niet verder gegaan met deze iteratie. |
| 12  Vervolg op iteratie 8 | I’m going to tip $300K for a better solution! Toevoegen aan de prompt van het LLM die antwoord geeft voor betere prestaties.  Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.849488 | 0.236529 | | 0.000000 | 0.000000 | 0.222113 | | 0.000000 | 0.997366 | 0.229966 | | 1.000000 | 0.987900 | 0.988076 | | 0.000000 | 0.000000 | 0.217593 | | 0.000000 | 0.000000 | 0.213153 | | 0.000000 | 0.904507 | 0.637953 | | 0.000000 | 0.949495 | 0.193357 | | 0.500000 | 0.951149 | 0.743648 | | 0.428571 | 0.000000 | 0.213886 |  * Gemiddelde faithfulness: 0.192857 * Gemiddelde answer\_relevancy: 0.56399 * Gemiddelde answer\_correctness: 0.389627 | | De answer correctness en faithfulness is omhoog gegaan. De answer\_relevancy is wel omlaag gegaan. Dat komt omdat hij bij de vragen die een answer relevancy van 0 hebben aangeeft dat hij de vraag niet kan beantwoorden op basis van de informatie. Dit is juist goed. Ik heb liever dat hij de vraag niet beantwoord dan iets verzint. Daarom wordt er verder gegaan met deze iteratie. |
| 13 | Formatting aan de prompt toevoegen. Dit zou moet zorgen voor betere prestaties.  Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.384615 | 0.0 | 0.197325 | | 0.000000 | 0.0 | 0.170425 | | 0.000000 | 0.0 | 0.161231 | | 0.500000 | 0.0 | 0.169981 | | 0.000000 | 0.0 | 0.171607 | | 0.000000 | 0.0 | 0.198783 | | 0.000000 | 0.0 | 0.546710 | | 0.066667 | 0.0 | 0.185965 | | 0.000000 | 0.0 | 0.219551 | | 0.000000 | 0.0 | 0.165522 |  * Gemiddelde faithfulness: 0.095128 * Gemiddelde answer\_relevancy: 0.0 * Gemiddelde answer\_correctness: 0.21871 | | Het model gaat zich hierdoor heel gek gedragen. |
| 14  Vervolg op iteratie 12 | Your task is to answer questions.  Toevoegen aan de prompt van het LLM die antwoord geeft voor betere prestaties.  Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.849488 | 0.236529 | | 0.0 | 0.933584 | 0.233269 | | 0.0 | 0.997366 | 0.229962 | | 1.0 | 0.987900 | 0.988076 | | 0.0 | 0.956245 | 0.235177 | | 1.0 | 0.761531 | 0.208429 | | 0.0 | 0.927260 | 0.198367 | | 0.0 | 0.949495 | 0.191648 | | 0.5 | 0.917839 | 0.747013 | | 0.0 | 0.937211 | 0.217756 |  * Gemiddelde faithfulness: 0.25 * Gemiddelde answer\_relevancy: 0.921792 * Gemiddelde answer\_correctness: 0.348623 | | De answer correctness is weer gedaald. De anser relevancy is gestegen dit komt omdat het model nu altijd antwoord geeft. |
| 15  Vervolg op iteratie 12 | You will be penalized for incorrect answers. Toevoegen aan de prompt van het LLM die antwoord geeft voor betere prestaties.  Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.00 | 0.907469 | 0.235096 | | 0.00 | 0.920896 | 0.226168 | | 0.00 | 0.997366 | 0.229959 | | 1.00 | 0.987900 | 0.986838 | | 0.00 | 0.000000 | 0.206647 | | 0.00 | 0.959228 | 0.229581 | | 0.00 | 0.931614 | 0.597714 | | 0.00 | 0.962928 | 0.197466 | | 0.00 | 0.000000 | 0.189350 | | 0.25 | 0.937581 | 0.217308 |  * Gemiddelde faithfulness: 0.125 * Gemiddelde answer\_relevancy: 0.760498 * Gemiddelde answer\_correctness: 0.331613 | | Dit werkt minder goed. |
| 16  Vervolg op iteratie 12 | Answer a question given in a natural, human-like manner. Toevoegen aan de prompt van het LLM die antwoord geeft voor betere prestaties.  Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.930616 | 0.235764 | | 0.000000 | 0.866734 | 0.233622 | | 0.200000 | 0.000000 | 0.204887 | | 0.500000 | 0.948049 | 0.722708 | | 0.000000 | 0.956245 | 0.231352 | | 0.000000 | 0.955562 | 0.222737 | | 0.142857 | 0.894724 | 0.205193 | | 0.000000 | 0.962928 | 0.194980 | | 0.250000 | 0.000000 | 0.530347 | | 0.125000 | 0.000000 | 0.188823 |  * Gemiddelde faithfulness: 0.121786 * Gemiddelde answer\_relevancy: 0.651486 * Gemiddelde answer\_correctness: 0.297041 | | Het model presteerd nu minder goed. |
| 17  Vervolg op iteratie 12 | Ensure that your answer is unbiased and avoids relying on stereotypes.  Toevoegen aan de prompt van het LLM die antwoord geeft voor betere prestaties.  Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.000000 | 0.230477 | | 0.000000 | 0.835264 | 0.223245 | | 1.000000 | 0.000000 | 0.203277 | | 1.000000 | 0.987900 | 0.988076 | | 0.000000 | 0.000000 | 0.181040 | | 0.400000 | 0.000000 | 0.202120 | | 0.000000 | 0.955876 | 0.203319 | | 0.000000 | 0.000000 | 0.193369 | | 0.333333 | 0.000000 | 0.587272 | | 0.000000 | 0.885581 | 0.207954 |  * Gemiddelde faithfulness: 0.273333 * Gemiddelde answer\_relevancy: 0.366462 * Gemiddelde answer\_correctness: 0.322015 | | Het model geeft slechter antwoord. |
| 18  Vervolg op iteratie 12 | think step by step  Toevoegen aan de prompt van het LLM die antwoord geeft voor betere prestaties.  Bron: *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.865706 | 0.236070 | | 0.000000 | 0.934709 | 0.225753 | | 0.500000 | 0.000000 | 0.210448 | | 1.000000 | 0.987900 | 0.988076 | | 1.000000 | 0.000000 | 0.212856 | | 0.000000 | 0.000000 | 0.722914 | | 0.000000 | 0.963276 | 0.207492 | | 0.000000 | 0.847194 | 0.199767 | | 0.666667 | 0.000000 | 0.534571 | | 0.157895 | 0.000000 | 0.216907 |  * Gemiddelde faithfulness: 0.332456 * Gemiddelde answer\_relevancy: 0.459879 * Gemiddelde answer\_correctness: 0.375485 | | Het model presteert minder goed. |
| 19  Vervolg op iteratie 12 | Iteraties 12, 14, 15 en 18 combineren voor een beter resultaat. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.792845 | 0.590820 | | 0.000000 | 0.936932 | 0.224271 | | 0.200000 | 0.000000 | 0.206136 | | 1.000000 | 0.987900 | 0.986838 | | 0.285714 | 0.000000 | 0.202294 | | 0.375000 | 0.000000 | 0.198315 | | 0.200000 | 0.872777 | 0.450384 | | 0.000000 | 0.000000 | 0.363263 | | 0.500000 | 0.951208 | 0.988354 | | 0.333333 | 0.000000 | 0.530782 |  * Gemiddelde faithfulness: 0.389405 * Gemiddelde answer\_relevancy: 0.454166 * Gemiddelde answer\_correctness: 0.474146 | | Het LLm geeft nu betere antwoorden. Alleen zijn deze antwoorden nu soms in het Engels. |
| 20 | Prompt naar het Nederlands vertalen. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.929284 | 0.731261 | | 0.000000 | 0.000000 | 0.222153 | | 0.000000 | 0.000000 | 0.207622 | | 0.166667 | 0.000000 | 0.433990 | | 0.500000 | 0.000000 | 0.596193 | | 0.000000 | 0.955562 | 0.606104 | | 0.125000 | 0.908757 | 0.211100 | | 0.166667 | 0.961789 | 0.187053 | | 0.500000 | 0.788111 | 0.714316 | | 0.058824 | 0.000000 | 0.219623 |  * Gemiddelde faithfulness: 0.151716 * Gemiddelde answer\_relevancy: 0.45435 * Gemiddelde answer\_correctness: 0.412942 | | De answer correctness en faitfulness zijn iets omlaag gegaan. |
| 21  Vervolg op iteratie 19 | Prompt Engels houden maar vragen of hij in het Nederlands kan antwoorden. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.928320 | 0.730391 | | 0.000000 | 0.991470 | 0.236437 | | 0.000000 | 0.997366 | 0.229959 | | 1.000000 | 0.987900 | 0.988076 | | 1.000000 | 0.000000 | 0.223100 | | 0.000000 | 0.000000 | 0.201904 | | 0.000000 | 0.927038 | 0.470613 | | 0.250000 | 0.962928 | 0.188658 | | 0.500000 | 0.840132 | 0.721074 | | 0.333333 | 0.945511 | 0.221575 |  * Gemiddelde faithfulness: 0.408333 * Gemiddelde answer\_relevancy: 0.758066 * Gemiddelde answer\_correctness: 0.421179 | | De answer correctness is iets lager maar toch wordt er voor gekozen om met deze verder te gaan want het model moet Nederlands praten. |
| 22  Vervolg op iteratie 21 | Onderdelen uit iteratie 19 ook toevoegen bij de prompt om de knowledge op te halen. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.000000 | 0.214960 | | 0.000000 | 0.957418 | 0.235720 | | 0.000000 | 0.878080 | 0.227678 | | 1.000000 | 0.879786 | 0.228183 | | 0.000000 | 0.000000 | 0.218837 | | 0.500000 | 0.000000 | 0.732663 | | 0.000000 | 0.000000 | 0.196468 | | 0.000000 | 0.000000 | 0.190921 | | 0.000000 | 0.000000 | 0.222460 | | 0.222222 | 0.000000 | 0.219623 |  * Gemiddelde faithfulness: 0.272222 * Gemiddelde answer\_relevancy: 0.271528 * Gemiddelde answer\_correctness: 0.268751 | |  |
| 23  Vervolg op iteratie 21 | Be thorough in your extraction of knowledg | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.929568 | 0.236564 | | 0.0 | 0.971482 | 0.236725 | | 0.0 | 0.000000 | 0.214315 | | 0.0 | 0.987900 | 0.986838 | | 0.0 | 0.967309 | 0.234168 | | 0.0 | 0.894578 | 0.224067 | | 0.0 | 0.957682 | 0.728817 | | 0.0 | 0.984250 | 0.198367 | | 0.5 | 0.978993 | 0.744532 | | 0.0 | 0.888948 | 0.391314 |  * Gemiddelde faithfulness: 0.05 * Gemiddelde answer\_relevancy: 0.856071 * Gemiddelde answer\_correctness: 0.419571 | |  |
| 24  Vervolg op iteratie 21 | Captions niet meenemen in de knowledge deze zijn vaak niet helemaal goed. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.849787 | 0.989019 | | 0.000000 | 1.000000 | 0.521868 | | 0.000000 | 0.000000 | 0.210886 | | 1.000000 | 1.000000 | 0.233282 | | 0.666667 | 0.000000 | 0.531697 | | 0.000000 | 0.000000 | 0.597102 | | 0.000000 | 0.923355 | 0.200467 | | 0.000000 | 0.962928 | 0.193467 | | 0.500000 | 0.894393 | 0.608496 | | 0.111111 | 0.000000 | 0.216973 |  * Gemiddelde faithfulness: 0.227778 * Gemiddelde answer\_relevancy: 0.563046 * Gemiddelde answer\_correctness: 0.430326 | | De anserwer correctness is een klein beetje omhooggegaan. |
| 25  Vervolg op iteratie 24 | Er zitten kleine foutjes in de transcriptie. Het model veranderen van int 8 naar float 32 voor beteren prestaties. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.806032 | 0.201662 | | 0.0 | 0.840577 | 0.521244 | | 0.0 | 0.885451 | 0.210736 | | 0.0 | 0.987891 | 0.986838 | | 0.0 | 0.000000 | 0.220396 | | 0.0 | 0.920933 | 0.227463 | | 0.0 | 0.957682 | 0.604713 | | 0.0 | 0.967721 | 0.574295 | | 0.0 | 0.000000 | 0.204949 | | 0.2 | 0.937211 | 0.221522 |  * Gemiddelde faithfulness: 0.02 * Gemiddelde answer\_relevancy: 0.73035 * Gemiddelde answer\_correctness: 0.397382 | | De transcriptie is vrijwel hetzelfde en niet veel beter dan eerst. De chunks zijn echter nu wel anders ingedeeld. Ik denk dat daar het verschil in prestaties door komt. |
| 25  Vervolg op iteratie 23 | Om het model minder te laten hallucineren. Answer strictly based on retrieved knowledge. Do not rely on prior knowledge or make assumptions beyond the retrieved information. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.928350 | 0.236019 | | 0.0 | 0.000000 | 0.223116 | | 0.5 | 0.891154 | 0.218855 | | 1.0 | 0.995967 | 0.233855 | | 0.5 | 0.000000 | 0.527495 | | 0.0 | 0.934905 | 0.735869 | | 0.0 | 0.000000 | 0.194369 | | 0.0 | 0.962928 | 0.191261 | | 0.0 | 0.947768 | 0.734892 | | 0.0 | 0.992883 | 0.593699 |  * Gemiddelde faithfulness: 0.2 * Gemiddelde answer\_relevancy: 0.665395 * Gemiddelde answer\_correctness: 0.388943 | | Het model presteert minder goed en lijkt niet minder te hallucineren. |
| 26  Vervolg op iteratie 23 | You are extracting structured knowledge from multimodal input data for use in a knowledge graph.  Deze zin toevoegen aan de prompt om betere triplets te genereren. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.928350 | 0.236019 | | 0.000000 | 0.959654 | 0.235737 | | 0.000000 | 0.930169 | 0.215171 | | 0.000000 | 0.842743 | 0.227001 | | 0.000000 | 0.000000 | 0.221249 | | 0.000000 | 0.950246 | 0.996514 | | 0.000000 | 0.853678 | 0.209507 | | 0.500000 | 0.841182 | 0.197775 | | 0.000000 | 0.896980 | 0.224961 | | 0.214286 | 0.000000 | 0.215935 |  * Gemiddelde faithfulness: 0.071429 * Gemiddelde answer\_relevancy: 0.7203 * Gemiddelde answer\_correctness: 0.297987 | | Dit werkt niet goed. |
| 27  Vervolg op iteratie 23 | Each triplet must represent a complete and logically valid fact, with no placeholders or vague references.  Deze zin toevoegen aan de prompt om betere triplets te genereren. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.000000 | 0.596631 | | 1.000000 | 0.000000 | 0.219127 | | 0.000000 | 0.997366 | 0.229959 | | 1.000000 | 0.987900 | 0.227862 | | 0.000000 | 0.792081 | 0.211769 | | 0.000000 | 0.951931 | 0.229965 | | 0.000000 | 0.977336 | 0.215801 | | 0.333333 | 0.891216 | 0.192963 | | 0.500000 | 0.978993 | 0.613280 | | 0.000000 | 0.932827 | 0.218754 |  * Gemiddelde faithfulness: 0.283333 * Gemiddelde answer\_relevancy: 0.750965 * Gemiddelde answer\_correctness: 0.295611 | | Dit werkt niet goed. |
| 28  Vervolg op iteratie 23 | Correct example: { "subject": "person", "predicate": "holds", "object": "cup" } Incorrect example: { "subject": "he", "predicate": "does", "object": "it" }  Zodat het model betere triplets maakt. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.996202 | 0.229620 | | 0.000000 | 0.000000 | 0.204440 | | 0.000000 | 0.866741 | 0.225447 | | 0.000000 | 0.875100 | 0.599825 | | 0.000000 | 0.792081 | 0.968288 | | 0.333333 | 0.767792 | 0.518919 | | 0.000000 | 0.000000 | 0.194369 | | 0.000000 | 0.962928 | 0.443993 | | 0.500000 | 0.978993 | 0.742672 | | 0.000000 | 0.000000 | 0.221215 |  * Gemiddelde faithfulness: 0.183333 * Gemiddelde answer\_relevancy: 0.623984 * Gemiddelde answer\_correctness: 0.434879 | | Dit heeft voor een kleine verbetering gezorgd. |
| 29  Vervolg op itertatie 28 | Itertatie 26 en 28 tegelijkertijd toevoegen. Iets vergelijkbaars eerder zorgt ook voor verbetering | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.938838 | 0.235685 | | 0.5 | 0.971482 | 0.238117 | | 1.0 | 0.000000 | 0.207818 | | 0.0 | 0.846782 | 0.205345 | | 0.5 | 0.956245 | 0.234835 | | 0.0 | 0.951084 | 0.232531 | | 0.0 | 0.893714 | 0.204878 | | 0.0 | 0.877133 | 0.194574 | | 1.0 | 0.961085 | 0.738385 | | 0.0 | 0.000000 | 0.593460 |  * Gemiddelde faithfulness: 0.3 * Gemiddelde answer\_relevancy: 0.739636 * Gemiddelde answer\_correctness: 0.308563 | | Dit zorgt niet voor een verbetering. |
| 30  Vervolg op itertatie 28 | Your goal is to extract clear, re-usable facts that could be added to a structured database or knowledge graph.  Deze zin toevoegen aan de prompt om betere triplets te genereren. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.833496 | 0.228392 | | 1.0 | 1.000000 | 0.221730 | | 0.5 | 0.891154 | 0.221898 | | 0.0 | 0.862383 | 0.231604 | | 0.0 | 0.000000 | 0.199088 | | 0.0 | 0.000000 | 0.191180 | | 0.0 | 0.913296 | 0.202255 | | 0.0 | 0.799343 | 0.198343 | | 0.5 | 0.961085 | 0.739326 | | 0.0 | 0.942030 | 0.219183 |  * Gemiddelde faithfulness: 0.2 * Gemiddelde answer\_relevancy: 0.720279 * Gemiddelde answer\_correctness: 0.2653 | | Werkt minder goed. |
| 31  Vervolg op itertatie 28 | Valuable knowledge consists of facts that are: (1) specific, (2) observable in the input, and (3) useful for understanding human-object interactions. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.841861 | 0.229713 | | 0.5 | 0.000000 | 0.222556 | | 0.0 | 0.874029 | 0.228403 | | 0.0 | 0.821592 | 0.977640 | | 0.0 | 0.000000 | 0.219779 | | 0.0 | 0.000000 | 0.203269 | | 0.0 | 0.000000 | 0.194369 | | 0.0 | 0.890710 | 0.495967 | | 0.0 | 0.928418 | 0.228825 | | 0.0 | 0.940970 | 0.216954 |  * Gemiddelde faithfulness: 0.05 * Gemiddelde answer\_relevancy: 0.529758 * Gemiddelde answer\_correctness: 0.321748 | | Nee werkt niet goed. |
| 32  Vervolg op itertatie 28 | Meer negatieve voorbeelden toevoegen: Incorrect example: { "subject": "person", "predicate": "is", "object": "there" } Incorrect example: { "subject": "object", "predicate": "exists", "object": "table" } | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.990671 | 0.232842 | | 0.000000 | 0.991470 | 0.236623 | | 0.000000 | 1.000000 | 0.226735 | | 0.000000 | 0.987913 | 0.225902 | | 0.500000 | 0.000000 | 0.218359 | | 0.000000 | 0.000000 | 0.224369 | | 0.333333 | 0.000000 | 0.198654 | | 0.000000 | 0.000000 | 0.199806 | | 0.500000 | 0.956302 | 0.736670 | | 0.142857 | 0.000000 | 0.216120 |  * Gemiddelde faithfulness: 0.147619 * Gemiddelde answer\_relevancy: 0.492636 * Gemiddelde answer\_correctness: 0.271608 | | Werkt niet goed. |
| 33  Vervolg op itertatie 28 | Do not include redundant or overly generic facts (e.g. 'person holds object') unless they add new, specific information. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.000000 | 0.521961 | | 0.000000 | 0.971482 | 0.232591 | | 0.000000 | 0.947436 | 0.211738 | | 0.500000 | 0.865743 | 0.213027 | | 0.000000 | 0.781811 | 0.206269 | | 0.000000 | 0.000000 | 0.191180 | | 0.000000 | 0.967207 | 0.202402 | | 0.000000 | 0.962928 | 0.573234 | | 0.333333 | 0.928418 | 0.228037 | | 0.333333 | 0.000000 | 0.222122 |  * Gemiddelde faithfulness: 0.116667 * Gemiddelde answer\_relevancy: 0.642502 * Gemiddelde answer\_correctness: 0.280256 | | Werkt niet goed. |
| 33  Vervolg op itertatie 28 | Replace vague terms like ‘he’, ‘she’, ‘it’ or ‘object’ with their specific referents from the input if possible. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.920711 | 0.233948 | | 0.0 | 0.978145 | 0.231322 | | 0.0 | 0.000000 | 0.207497 | | 0.5 | 0.813572 | 0.734525 | | 0.0 | 0.956245 | 0.998890 | | 0.0 | 0.000000 | 0.215293 | | 0.0 | 0.894561 | 0.202961 | | 0.0 | 0.993021 | 0.194645 | | 0.5 | 0.894393 | 0.608582 | | 0.4 | 0.913582 | 0.220675 |  * Gemiddelde faithfulness: 0.14 * Gemiddelde answer\_relevancy: 0.736423 * Gemiddelde answer\_correctness: 0.384834 | | Werkt niet goed. |
| 34  Vervolg op itertatie 28 | Use strong, descriptive predicates (e.g. 'grabs', 'places', 'hands over') instead of vague verbs like 'does', 'makes', 'moves' | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.0 | 0.000000 | 0.222189 | | 0.5 | 0.971482 | 0.233284 | | 0.0 | 1.000000 | 0.228088 | | 0.0 | 0.987900 | 0.228234 | | 0.0 | 0.869936 | 0.228321 | | 0.5 | 0.822536 | 0.225668 | | 0.0 | 0.880125 | 0.644579 | | 0.0 | 0.962928 | 0.193931 | | 0.5 | 0.963712 | 0.737643 | | 0.0 | 0.937458 | 0.219395 |  * Gemiddelde faithfulness: 0.25 * Gemiddelde answer\_relevancy: 0.839608 * Gemiddelde answer\_correctness: 0.316133 | | Werkt niet goed. |
| 35  Vervolg op itertatie 28 | Triplets should reflect interactions or changes involving objects and agents, not static descriptions. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0000 | 0.980629 | 0.233484 | | 0.0000 | 0.000000 | 0.223336 | | 0.0000 | 0.964910 | 0.221797 | | 0.0000 | 0.821470 | 0.224943 | | 0.0000 | 0.938824 | 0.227005 | | 0.5000 | 0.000000 | 0.209298 | | 0.0000 | 0.822639 | 0.208501 | | 0.0000 | 0.000000 | 0.195993 | | 0.5000 | 0.978993 | 0.227190 | | 0.3125 | 0.933892 | 0.215231 |  * Gemiddelde faithfulness: 0.13125 * Gemiddelde answer\_relevancy: 0.644136 * Gemiddelde answer\_correctness: 0.218678 | | Werkt niet goed. |
| 36  Vervolg op itertatie 28 | Ignore conversational filler, emotional expressions, or unrelated background details. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.880288 | 0.218822 | | 0.0 | 0.971488 | 0.238856 | | 0.0 | 0.000000 | 0.205129 | | 0.0 | 0.924250 | 0.225011 | | 0.0 | 0.956245 | 0.234851 | | 0.0 | 0.000000 | 0.228393 | | 0.0 | 0.957682 | 0.205562 | | 0.0 | 0.962928 | 0.19167 | | 0.0 | 0.966190 | 0.230863 | | 0.5 | 0.000000 | 0.721170 |  * Gemiddelde faithfulness: 0.05 * Gemiddelde answer\_relevancy: 0.661907 * Gemiddelde answer\_correctness: 0.270033 | | Werkt niet goed. |
| 37  Vervolg op itertatie 28 | "First, identify the most significant human-object or object-object interactions in the input.  Then, rewrite these interactions as S-P-O triplets using specific nouns and verbs." | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.873098 | 0.206607 | | 0.500000 | 0.971482 | 0.232841 | | 0.000000 | 0.874029 | 0.225655 | | 0.000000 | 1.000000 | 0.232275 | | 0.000000 | 0.731847 | 0.964528 | | 0.000000 | 0.889764 | 0.989971 | | 0.000000 | 0.886657 | 0.197892 | | 0.000000 | 0.962928 | 0.193540 | | 0.272727 | 0.000000 | 0.226514 | | 0.200000 | 0.937211 | 0.219974 |  * Gemiddelde faithfulness: 0.197273 * Gemiddelde answer\_relevancy: 0.812702 * Gemiddelde answer\_correctness: 0.36898 | | Werkt niet goed. |
| 38  Vervolg op itertatie 28 | Rank potential triplets internally and include only the ones that are: (1) most specific, (2) most causally relevant, and (3) most observable | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.819249 | 0.216863 | | 0.0 | 0.991471 | 0.236623 | | 0.0 | 0.000000 | 0.209066 | | 0.6 | 0.000000 | 0.219445 | | 0.0 | 0.000000 | 0.218210 | | 0.0 | 0.000000 | 0.191182 | | 0.0 | 0.957682 | 0.212216 | | 0.0 | 0.962928 | 0.195810 | | 0.5 | 0.749374 | 0.719219 | | 0.0 | 0.000000 | 0.217032 |  * Gemiddelde faithfulness: 0.11 * Gemiddelde answer\_relevancy: 0.44807 * Gemiddelde answer\_correctness: 0.263566 | |  |
| 39  Vervolg op itertatie 28 | Review each sentence and ask: ‘Is this important enough to be stored as knowledge for future use?’ Only then extract. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.918219 | 0.228462 | | 0.0 | 0.971482 | 0.231843 | | 0.0 | 0.997366 | 0.228688 | | 0.0 | 0.862385 | 0.231604 | | 0.0 | 0.000000 | 0.214314 | | 0.0 | 0.889764 | 0.221133 | | 0.0 | 0.891692 | 0.211462 | | 0.0 | 0.962928 | 0.191723 | | 0.5 | 0.000000 | 0.217816 | | 0.0 | 0.000000 | 0.216495 |  * Gemiddelde faithfulness: 0.05 * Gemiddelde answer\_relevancy: 0.649383 * Gemiddelde answer\_correctness: 0.219354 | | Werkt niet goed. |
| 40  Vervolg op itertatie 28 | "The triplets are used to answer questions about videos. Extract only what is important to answer the questions." | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.000000 | 0.217044 | | 0.5 | 0.000000 | 0.526029 | | 0.0 | 0.000000 | 0.215796 | | 0.0 | 0.807986 | 0.204611 | | 0.0 | 0.000000 | 0.214649 | | 0.0 | 0.943477 | 0.233413 | | 0.0 | 0.910228 | 0.194380 | | 0.0 | 0.980546 | 0.191373 | | 0.0 | 0.938633 | 0.229558 | | 0.0 | 0.933892 | 0.217781 |  * Gemiddelde faithfulness: 0.05 * Gemiddelde answer\_relevancy: 0.551476 * Gemiddelde answer\_correctness: 0.244463 | | Werkt niet goed. |
| 41  Vervolg op itertatie 28 | Prioriteit in de prompt toevoegen aan de objecten waar de meeste acties mee uitgevoerd worden of het meeste aangeraakt worden. Bijvoorbeeld: Focus objects in this chunk: Kabel, Schroevendraaier, Momentsleutel | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.00 | 0.930849 | 0.23078 | | 0.00 | 0.898795 | 0.213325 | | 0.00 | 0.000000 | 0.217329 | | 0.00 | 0.858188 | 0.225834 | | 0.00 | 0.956245 | 0.241866 | | 0.00 | 0.955562 | 0.240840 | | 0.00 | 0.000000 | 0.195960 | | 0.00 | 0.962928 | 0.198295 | | 0.00 | 0.908626 | 0.222734 | | 0.25 | 0.942030 | 0.219791 |  * Gemiddelde faithfulness: 0.125 * Gemiddelde answer\_relevancy: 0.741322 * Gemiddelde answer\_correctness: 0.220676 | | Werkt niet goed. |
| 42  Vervolg op itertatie 28 | After extracting triplets, re-check if each one is logically consistent and specific. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.931816 | 0.236147 | | 0.0 | 0.804700 | 0.194651 | | 0.0 | 1.000000 | 0.226733 | | 0.0 | 0.862383 | 0.231604 | | 0.0 | 0.000000 | 0.657849 | | 0.0 | 0.889764 | 0.221145 | | 0.0 | 0.992965 | 0.202297 | | 0.0 | 0.962928 | 0.192993 | | 0.5 | 0.963055 | 0.738142 | | 0.0 | 0.000000 | 0.216446 |  * Gemiddelde faithfulness: 0.05 * Gemiddelde answer\_relevancy: 0.740761 * Gemiddelde answer\_correctness: 0.311801 | |  |
| 43  Vervolg op itertatie 28 | If you can deduce a clear causal relation (e.g., inserting a part requires tool X), include this as a separate triple. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.980629 | 0.232035 | | 0.000000 | 0.991470 | 0.238864 | | 0.000000 | 0.904369 | 0.225837 | | 0.000000 | 0.987900 | 0.233855 | | 0.000000 | 0.000000 | 0.211362 | | 0.000000 | 0.000000 | 0.191180 | | 0.000000 | 0.000000 | 0.196558 | | 0.000000 | 0.000000 | 0.196146 | | 0.000000 | 0.899032 | 0.225845 | | 0.307692 | 0.000000 | 0.215375 |  * Gemiddelde faithfulness: 0.030769 * Gemiddelde answer\_relevancy: 0.47634 * Gemiddelde answer\_correctness: 0.216706 | | Werkt niet goed. |
| 44  Vervolg op itertatie 28 | After generating the triplets, review each one and remove any that are not clearly supported by the input. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.000000 | 0.220228 | | 0.000000 | 0.962541 | 0.229267 | | 0.000000 | 0.935795 | 0.222880 | | 0.500000 | 0.987900 | 0.230372 | | 0.000000 | 0.000000 | 0.719373 | | 0.000000 | 0.000000 | 0.191180 | | 0.000000 | 0.936940 | 0.517408 | | 0.000000 | 0.962928 | 0.199972 | | 0.800000 | 0.000000 | 0.427554 | | 0.461538 | 0.937581 | 0.216702 |  * Gemiddelde faithfulness: 0.176154 * Gemiddelde answer\_relevancy: 0.572368 * Gemiddelde answer\_correctness: 0.317494 | |  |
| 45  Vervolg op itertatie 28 | Hele nieuwe prompt: You are an expert in transforming multimodal scene data into structured knowledge triples (subject, predicate, object). You will receive:  1. A natural language transcription of audio  2. A visual scene description  3. A list of detected objects  Your task:  - Combine these sources to extract \*\*clear, grounded, and non-redundant\*\* knowledge triples.  - Use \*\*specific terminology\*\* found in the input.  - Only include \*\*actions or relations directly supported\*\* by the data.  DO NOT:  - Include vague terms like “object”, “person”, “thing”  - Infer physical states that are not directly visible or stated  - Repeat the same relation using synonyms | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.932970 | 0.234884 | | 0.500000 | 0.910961 | 0.221098 | | 0.000000 | 0.000000 | 0.216381 | | 0.000000 | 0.987900 | 0.227047 | | 0.500000 | 0.000000 | 0.601476 | | 0.000000 | 0.955562 | 0.242562 | | 0.000000 | 0.928437 | 0.197200 | | 0.000000 | 0.967721 | 0.200300 | | 0.500000 | 0.938833 | 0.745301 | | 0.333333 | 0.933892 | 0.216901 |  * Gemiddelde faithfulness: 0.183333 * Gemiddelde answer\_relevancy: 0.755628 * Gemiddelde answer\_correctness: 0.310315 | |  |
| 46  Vervolg op itertatie 28 | You are a knowledge extraction assistant. \n"  "Given a technical sentence, extract all factual statements it contains in the form of subject–predicate–object triples. \n"  "Each triple must correspond to an explicit fact stated in the input text — do not infer or assume. \n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.873098 | 0.219789 | | 0.0 | 0.884511 | 0.225954 | | 0.0 | 0.997366 | 0.229959 | | 0.0 | 0.988074 | 0.226679 | | 0.0 | 0.000000 | 0.223595 | | 0.0 | 0.000000 | 0.216343 | | 0.0 | 0.935313 | 0.209627 | | 0.0 | 0.000000 | 0.190622 | | 0.0 | 0.000000 | 0.224802 | | 0.0 | 0.000000 | 0.434925 |  * Gemiddelde faithfulness: 0.0 * Gemiddelde answer\_relevancy: 0.467836 * Gemiddelde answer\_correctness: 0.240229 | |  |
| 47  Vervolg op itertatie 28 | "You are an information extraction engine. \n"                          "Your task is to extract all relationships (triples). \n"                          "You will work in three steps: \n"                          "1. Identify all noun phrases that can act as subjects or objects. \n"                          "2. Identify all verbs or relational phrases connecting them. \n"                          "3. Construct one triple per relationship found. \n"                          "Rules: \n"                          "- Use only information stated \*explicitly\* in the input. \n"                          "- Do not interpret, infer, or paraphrase. \n"                          "- If multiple triples exist, return them all. \n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.0 | 0.928680 | 0.235082 | | 0.0 | 0.892437 | 0.236216 | | 1.0 | 0.997366 | 0.229959 | | 0.0 | 0.987900 | 0.226262 | | 0.0 | 0.956245 | 0.233556 | | 0.5 | 0.893014 | 0.228556 | | 0.0 | 0.000000 | 0.186723 | | 0.0 | 0.939123 | 0.191191 | | 0.5 | 0.978993 | 0.737992 | | 0.0 | 0.941297 | 0.218221 |  * Gemiddelde faithfulness: 0.3 * Gemiddelde answer\_relevancy: 0.851505 * Gemiddelde answer\_correctness: 0.272376 |  | |
| 48  Vervolg op itertatie 28 | "First, list all noun phrases (NPs) in the input sentence. \n"  "Then, for each NP, find all direct relations it has with other NPs in the form of [subject, relation, object]. \n" "Return only explicit relations. \n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.873098 | 0.220333 | | 0.000000 | 0.971482 | 0.229982 | | 0.000000 | 0.921021 | 0.217028 | | 0.500000 | 0.000000 | 0.717156 | | 0.500000 | 0.000000 | 0.389830 | | 0.000000 | 0.000000 | 0.191180 | | 0.000000 | 0.875676 | 0.204049 | | 0.000000 | 0.860255 | 0.195879 | | 0.333333 | 0.865441 | 0.232904 | | 0.333333 | 0.000000 | 0.510736 |  * Gemiddelde faithfulness: 0.166667 * Gemiddelde answer\_relevancy: 0.536697 * Gemiddelde answer\_correctness: 0.310908 |  | |
| 49  Vervolg op itertatie 28 | "You are a knowledge extraction engine.\n"  "From the input, extract all factual subject–predicate–object triplets that represent meaningful relationships between entities, actions, or objects.\n"  "Be exhaustive: extract as many valid triplets as the input allows, but do not invent any.\n"  "Only extract triplets that are clear and semantically valid.\n"  "Use generic terms like 'person' or 'object' when references are ambiguous.\n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.2 | 0.000000 | 0.225379 | | 0.5 | 0.908798 | 0.231672 | | 0.0 | 0.989400 | 0.218307 | | 0.0 | 0.000000 | 0.195302 | | 0.0 | 0.000000 | 0.220244 | | 0.0 | 0.000000 | 0.191180 | | 0.0 | 0.937608 | 0.210183 | | 0.0 | 0.962928 | 0.961059 | | 0.5 | 0.854563 | 0.227621 | | 1.0 | 0.000000 | 0.211410 |  * Gemiddelde faithfulness: 0.22 * Gemiddelde answer\_relevancy: 0.46533 * Gemiddelde answer\_correctness: 0.289236 |  | |
| 50  Vervolg op itertatie 28 | Lange few-shot prompt  "You are a knowledge extractor. Extract subject-predicate-object knowledge triplets ""from the given input. Each triplet must contain:\n"  "- A concrete subject (e.g., 'person', not 'he')\n"  "- A meaningful predicate (action or relation)\n"  "- A concrete object (e.g., 'cup', not 'it')\n\n"  "Return ONLY valid triplets in the following JSON schema:\n"    "{\n""  \"valuable\_knowledge\": [\n""    { \"subject\": \"...\", \"predicate\": \"...\", \"object\": \"...\" }\n""  ]\n""}\n" "If no knowledge can be extracted, return:\n""{\"valuable\_knowledge\": []}\n\n""Examples:\n\n"  "Input:\n""\"The person grabs a knife and cuts a cucumber on the cutting board.\"\n" "Output:\n""{\n""  \"valuable\_knowledge\": [\n"{ \"subject\": \"person\", \"predicate\": \"grabs\", \"object\": \"knife\" },\n""    { \"subject\": \"person\", \"predicate\": \"cuts\", \"object\": \"cucumber\" },\n"  "    { \"subject\": \"cucumber\", \"predicate\": \"is on\", \"object\": \"cutting board\" }\n""  ]\n""}\n\n"Input:\n""\"The individual picks up a red cup and moves it to the sink.\"\n""Output:\n""{\n""  \"valuable\_knowledge\": [\n"                          "    { \"subject\": \"person\", \"predicate\": \"picks up\", \"object\": \"red cup\" },\n"                          "    { \"subject\": \"person\", \"predicate\": \"moves\", \"object\": \"red cup\" },\n""    { \"subject\": \"red cup\", \"predicate\": \"is moved to\", \"object\": \"sink\" }\n""  ]\n""}\n\n""Now extract knowledge from the next input:" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 1.000000 | 0.943315 | 0.221540 | | 0.500000 | 0.883340 | 0.223668 | | 0.333333 | 0.000000 | 0.207618 | | 1.000000 | 0.898444 | 0.229771 | | 0.000000 | 0.000000 | 0.222158 | | 0.000000 | 0.000000 | 0.191180 | | 0.000000 | 0.969103 | 0.201134 | | 0.000000 | 0.000000 | 0.192053 | | 0.500000 | 0.978993 | 0.738916 | | 0.133333 | 0.000000 | 0.216800 |  * Gemiddelde faithfulness: 0.346667 * Gemiddelde answer\_relevancy: 0.467319 * Gemiddelde answer\_correctness: 0.264484 |  | |
| 51  Vervolg op itertatie 28 | "You are a knowledge extraction engine. Your task is to extract only meaningful and generalizable subject-predicate-object triplets. "   "Only include triplets that express a clear physical or functional relationship, interaction, or action relevant to the real world. "   "Avoid personal pronouns or vague verbs. Always normalize subjects to concrete entities like 'person', 'knife', or 'cup'.\n\n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.500000 | 0.785762 | 0.209158 | | 0.000000 | 0.984807 | 0.235480 | | 0.000000 | 0.874029 | 0.225529 | | 0.000000 | 0.950146 | 0.221926 | | 0.000000 | 0.956245 | 0.235091 | | 0.000000 | 0.876254 | 0.221260 | | 0.428571 | 0.000000 | 0.213702 | | 0.000000 | 0.962928 | 0.194339 | | 0.250000 | 0.941075 | 0.486385 | | 0.375000 | 0.000000 | 0.217070 |  * Gemiddelde faithfulness: 0.155357 * Gemiddelde answer\_relevancy: 0.733125 * Gemiddelde answer\_correctness: 0.245994 |  | |
| 52  Vervolg op itertatie 28 | "Use normalized and concrete terms (e.g., 'person', 'cup') for subjects and objects. Avoid vague verbs or pronouns.\n"  Deze zin toegevoegd. | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.500000 | 1.000000 | 0.234062 | | 0.500000 | 0.971482 | 0.233078 | | 0.000000 | 0.980823 | 0.216404 | | 1.000000 | 0.867508 | 0.231003 | | 0.285714 | 0.000000 | 0.438784 | | 0.000000 | 0.953604 | 0.239703 | | 0.000000 | 0.889684 | 0.206357 | | 0.000000 | 0.799343 | 0.199735 | | 0.500000 | 0.978993 | 0.234735 | | 0.250000 | 0.945511 | 0.221613 |  * Gemiddelde faithfulness: 0.303571 * Gemiddelde answer\_relevancy: 0.838695 * Gemiddelde answer\_correctness: 0.245547 |  | |
| 53  Vervolg op itertatie 28 | "You are a knowledge extraction agent. Your task is to extract as many distinct, factual subject-predicate-object triplets as possible from the provided transcript, object-action pairs, and hand-object interactions.\n"  "Each triplet should express a complete, concrete action or relation, using general terms like 'person' instead of pronouns, and specific verbs (e.g., 'holds', 'opens').\n"  "Only include triplets that are grounded in the input and avoid vague or inferred information.\n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.000000 | 0.921087 | 0.235969 | | 0.000000 | 0.000000 | 0.221358 | | 0.000000 | 0.000000 | 0.516205 | | 0.000000 | 0.000000 | 0.206863 | | 0.000000 | 0.000000 | 0.199419 | | 0.000000 | 0.000000 | 0.205968 | | 0.000000 | 0.922990 | 0.211126 | | 0.333333 | 0.000000 | 0.193930 | | 0.500000 | 0.878162 | 0.541812 | | 0.166667 | 0.945511 | 0.222767 |  * Gemiddelde faithfulness: 0.1 * Gemiddelde answer\_relevancy: 0.366775 * Gemiddelde answer\_correctness: 0.275542 |  | |
| 54  Vervolg op itertatie 28 | "You are a knowledge extractor. Extract subject-predicate-object knowledge triplets from all provided information: transcript, linked object-action pairs, and hand-object interactions.\n"  "Use generic subjects like 'person' and concrete predicates like 'grabs', 'places', 'opens'.\n"  "Each triplet must reflect an explicit, observable interaction or relation.\n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.923627 | 0.228035 | | 0.0 | 0.991470 | 0.236006 | | 0.0 | 0.000000 | 0.213244 | | 0.0 | 0.987900 | 0.228509 | | 0.0 | 0.956245 | 0.235285 | | 0.0 | 0.889764 | 0.221506 | | 0.0 | 0.844623 | 0.212985 | | 0.0 | 0.899316 | 0.193596 | | 0.0 | 0.978993 | 0.235477 | | 0.0 | 0.946719 | 0.217205 |  * Gemiddelde faithfulness: 0.0 * Gemiddelde answer\_relevancy: 0.841866 * Gemiddelde answer\_correctness: 0.222185 |  | |
| 55  Vervolg op itertatie 28 | "You are a knowledge extractor. Your job is to extract all relevant and specific subject-predicate-object triplets based on the transcript, linked object-action pairs, and hand-object interactions.\n"  "Use general nouns like 'person' instead of pronouns, and avoid vague verbs like 'does' or 'moves'. Prefer specific verbs that describe physical actions.\n"  "Only include triplets that are clearly supported by the input.\n" | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.0 | 0.000000 | 0.703857 | | 0.0 | 0.960583 | 0.241455 | | 0.0 | 0.997366 | 0.229486 | | 0.0 | 0.987900 | 0.230734 | | 0.0 | 0.792081 | 0.214213 | | 0.0 | 0.889764 | 0.223117 | | 0.0 | 0.912915 | 0.765465 | | 0.0 | 0.962928 | 0.192467 | | 0.0 | 0.925010 | 0.229699 | | 0.0 | 0.000000 | 0.224207 |  * Gemiddelde faithfulness: 0.0 * Gemiddelde answer\_relevancy: 0.742855 * Gemiddelde answer\_correctness: 0.32547 |  | |
| 56  Vervolg op itertatie 28 | Triplets achteraf laten controleren door LLM |  |  | |
| 57  Vervolg op itertatie 28 | Nieuwe prompt ontwikkeld om kennis te extraheren. Waarbij het model is gevraagd wordt om een hypothese op te stellen over wat er gebeurt in de scene. Vervolgens wordt er gevraagd om te redeneren over deze hypothese en wat er onzeker aan is. Hierbij wordt er ook gevraagd om een lijst met onzekerheden om te stellen. Vervolgens worden er triplets opgesteld.  Dit is de prompt:  "You are a reasoning entity. Analyze the following observations from a scene. "  "First, hypothesize what is happening, then reflect on what is uncertain. "   "Return a JSON with keys: hypothesis, reasoning, doubts (list), knowledge (triplets). "  "Each triplet has subject, predicate, object." | |  |  |  | | --- | --- | --- | | **faithfulness** | **Answer relevancy** | **Answer correctness** | | 0.00 | 0.000000 | 0.190122 | | 0.00 | 0.826718 | 0.210694 | | 1.00 | 0.997366 | 0.228960 | | 1.00 | 0.948213 | 0.231032 | | 0.00 | 0.956245 | 0.998139 | | 0.00 | 0.889764 | 0.989216 | | 0.00 | 0.957685 | 0.206783 | | 0.00 | 0.962928 | 0.576628 | | 0.00 | 0.960928 | 0.738005 | | 0.25 | 0.000000 | 0.820401 |  * Gemiddelde faithfulness: 0.225 * Gemiddelde answer\_relevancy: 0.749985 * Gemiddelde answer\_correctness: 0.518998 | De nieuwe methode werkt beter. Dit komt vooral doordat de triplets nu meer correcte informatie bevatten er minder onjuiste informatie inzit. | |
| 58  Vervolg op itertatie 57 | "You are a machine of pure inference. Your task is to extract structured truth from unstructured observation. Treat all input as forensic evidence."                  "Proceed in four strict phases:"                  "(1) HYPOTHESIS — State, in one precise sentence, what is most likely happening. Be decisive."                  "(2) REASONING — Justify your hypothesis through a clear, deductive chain of logic. Each step must follow from evidence."                  "(3) DOUBTS — List every unknown, ambiguity, or alternative interpretation. Reveal the edges of your certainty."                  "(4) KNOWLEDGE — Extract only hard facts. Output them as [subject, predicate, object] triplets."                  "Return a JSON object with keys: 'hypothesis', 'reasoning', 'doubts' (list), and 'knowledge' (list of triplets)."                  "Assume nothing. Invent nothing. Only infer what the data demands." |  |  | |