

Issue 26(4), 2026
ISSN: 1567-7141
DOI: 10.59490/ejtir.2026.26.4.8374

TU Delft | OPEN Publishing

Research Article

Capabilities of Gemini, GPT, and DeepSeek for transportation in Dutch: Assessing domain-specific and underrepresented language challenges

Thi M.D. Tran [1,*], Davy Janssens [1], Geert Wets [1], Tom Brijs [1], Burcu Can [2], Jan Vuurstaek [1], Lien Aerts[1], Wim Ectors [1]

[1] UHasselt, Transportation Research Institute, Martelarenlaan 42, 3500 Hasselt, Belgium
[2] Computing Science and Mathematics, University of Stirling, UK
* corresponding author: thi.tran@uhasselt.be

Keywords
transportation
large language model
underrepresented language
domain-specific task
GPT
gemini
DeepSeek

Publishing history
Submitted: 17 August 2025
Revised date(s): 11 March 2026, 14 July 2026
Accepted: 14 July 2026
Published: 30 October 2026

Cite as
Tran, T. M., Janssens, D., Wets, G., Brijs, T., Can, B., Vuurstaek, J., … Ectors, W. (2026). Capabilities of Gemini, GPT, and DeepSeek for transportation in Dutch: Assessing domain-specific and underrepresented language challenges. *European Journal of Transport and Infrastructure Research, 26(4).* https://doi.org/10.59490/ejtir.2026.26.4.8374

©2026 Authors. Published by TU Delft OPEN Publishing on behalf of the authors. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Abstract

Large Language Models (LLMs) have showcased remarkable progress in language comprehension and interactive decision-making, primarily in English and across general tasks. However, their capabilities in domain-specific reasoning and underrepresented languages remain underexplored. This study investigates the performance of LLMs in the Dutch language within the transportation domain, with a focus on cross-lingual generation and multimodal input processing. It aims to understand LLM's model behavior in the transportation domain to help select appropriate foundation models for further fine-tuning on specific tasks when training from scratch is not feasible due to data availability or high costs. We evaluate the performance of nine different LLMs from three prominent open- and closed-source LLM families: Gemini, GPT, and DeepSeek, using 972 questions distributed across six datasets. These datasets vary in complexity, input modality, and in English or Dutch to comprehensively assess the models. Our findings show that while LLMs generally perform better in English, several models demonstrate strong cross-lingual capabilities in Dutch. The DeepSeek R1 and GPT-4o models perform relatively well in both languages, while Gemini 1.5-Pro struggles with Dutch content. Additionally, LLMs exhibit a solid understanding of the context of transportation and regulation by effectively interpreting both information from images and text to answer domain-specific questions. However, they still lag behind the performance of elementary children in comprehensive transportation-related tasks, suggesting that LLM reasoning under zero-shot conditions may be limited. This study underscores the potential of LLMs in domain-specific applications and underrepresented languages and supports informed selection of foundation models for fine-tuning in these context settings.

1 Introduction

Generative large language models (LLMs) are designed to produce digital content, including text, code, images, video, and music, from input instructions. Although the underlying technology has been developed over many years, tools such as ChatGPT (OpenAI, 2022a) and DALL-E 2 (OpenAI, 2022b) have recently brought this technology to the forefront (Dwivedi et al., 2023), capturing public interest and stimulating further research.

In recent years, LLMs have rapidly evolved, scaling up in size, training data, and capabilities. GPT model series expanded from GPT-1 (117M parameters, 4.5 GB text) to GPT-3 (175B parameters, 570 GB text) and later GPT-4 (reported 100T parameters) (De Angelis et al., 2023; Ray, 2023). Google's Gemini series, including 1.0-Pro, 1.5-Pro, 1.5-Flash, 2.0-Flash, and experimental 2.5-Pro, introduced multimodal and reasoning improvements (Gemini Team, 2023; Pichai & Hassabis, 2024). OpenAI's GPT-4o and GPT-o1 expanded multilingual and multimodal performance, (OpenAI, 2024a) while DeepSeek R1 demonstrated that open-source models can achieve high performance with 671B parameters (DeepSeek API Docs, 2025). These developments highlight the rapid progress and growing potential of LLMs for research and domain-specific applications.

Alongside their technical evolution, LLMs have demonstrated impressive general-purpose capabilities, especially in high-resource languages, like English. They perform well on natural language processing tasks involving reasoning, summarization, information retrieval, and complex decision-making (Chang et al., 2024; Fan et al., 2024). However, their effectiveness tends to decline in underrepresented languages, where both the quantity and quality of training data are limited (Ahuja et al., 2023; Bang et al., 2023; Buscaldi & Rosso, 2023; Lai et al., 2023; Muller et al., 2021; Nazi et al., 2025). This disparity is especially important to address in specialized domains like transportation, where domain-specific terminology and contextual understanding are crucial.

Recent transportation research has begun investigating how large language models (LLMs) can support tasks such as traffic management, safety analysis, behavior prediction, and multimodal data integration (Zhang et al., 2025; Yan & Li, 2025; Guo et al., 2024; Mo et al., 2023; Li et al., 2024; Zhang et al., 2023; Liang et al., 2024). However, most existing studies focus on English or rely on a single base model, providing limited insight into cross-lingual capabilities and strategies for foundation model selection. Consequently, comprehensive evaluations of LLM performance in underrepresented languages, such as Dutch, remain scarce within transportation contexts. In this work, we systematically examine the cross-lingual and multimodal performance of state-of-the-art foundation models on transportation-related multiple-choice questions across six diverse datasets, with content and instruction prompts in either Dutch or English. The datasets include both text-only and text-image questions, reflecting the inherently multimodal nature of transportation data. These analyses provide guidance for informed LLM selection for domain-specific applications in European transportation, with particular emphasis on Dutch as a localized and underrepresented language.

Contributions

- Comprehensive evaluation of state-of-the-art LLMs (Gemini, GPT, DeepSeek) on transportation tasks in Dutch, providing insights to guide informed foundation model selection.
- Analysis of instructional prompting, highlighting model-specific strengths, weaknesses, and sensitivities to the language of instruction.
- Introduction of a specialized multimodal test dataset (972 questions across six datasets) that reflects the natural, multimodal characteristics of transportation and covers a wide range of traffic rules and road safety scenarios.

The remainder of this paper is structured as follows. Section 2 presents a review of related work, focusing on LLMs' multilingual performance, their potential in transportation, and the limited

coverage in prior studies. Section 3 details the methodology, including experimental setup, dataset construction, and evaluation method. Section 4 reports and analyzes the experimental results. Section 5 discusses key findings and their implications. Finally, section 6 concludes the paper, outlines the limitations, and provides directions for future research.

2 Related work

The evolution of language models can be traced back to the 1950s, from early statistical modeling techniques such as Hidden Markov Models (Knill & Young, 1997) and Gaussian Mixture Models (Reynolds, 2009) to neural language models, with the development of deep neural networks. However, a significant breakthrough occurred with the introduction of the transformer architecture and the self-attention mechanism in 2017 by Vaswani et al. (2017). This architecture has been effectively used in recent LLMs, allowing practical training on large datasets and the creation of scalable pre-trained models that can be fine-tuned for specific tasks and domains (Cao et al., 2025), leading to the state-of-the-art performance we see today.

In recent years, LLMs have shown growing potential in the transportation domain, enabling a range of applications from traffic management and planning to safety enhancement by effectively handling spatio-temporal data, performing complex analyses, forecasting behavior, and integrating multimodal traffic information (Zhang et al., 2025; Yan & Li, 2025; Guo et al., 2024; Mo et al., 2023; Li et al., 2024; Zhang et al., 2023; Liang et al., 2024). Several studies highlight their applicability: for instance, GPT-3.5 has been applied with k-shot prompting to predict transportation mode choices by incorporating domain-specific knowledge (Zhang et al., 2025). Guo et al. (2024) applied supervised fine-tuning to LLaMA 2 for explainable traffic flow prediction, translating multi-modal traffic data into natural language descriptions. This approach highlights the potential of LLMs in tackling spatio-temporal tasks within the transportation domain. Similarly, Mo et al. (2023) used GPT-3.5 with zero-shot prompting to predict travel behavior, focusing on mode choice and trip purpose, while improving interpretability by clarifying the reasoning behind predictions.

In autonomous driving, Li et al. (2024) introduced UnstrPrompt, a dataset of 6,000 paired language descriptions and image samples, generated using GPT-3.5, to support driving tasks in unstructured environments. Zhang et al. (2023) proposed TrafficGPT, a system combining GPT-3.5 Turbo with chain-of-thought prompting to process traffic data, facilitate human decision-making through dialogue, and enable interactive feedback for urban traffic management. A recent study by Liang et al. (2024) applied a chain-of-thought strategy to GPT-4 to predict human mobility during public events, drawing on online sources such as ticketing platforms and social media.

Most studies employ adaptation techniques to tailor general-purpose models for improved performance on domain-specific tasks. However, they typically rely on a single foundation model when applying these techniques, often without providing comparisons or clear justifications for the model choice. Alizadeh et al. (2024), in their study on text annotation tasks involving fine-tuning of LLMs, found that LLaMA, which initially lagged behind the base model GPT-3.5, surpassed the base GPT-3.5 after fine-tuning, but remained behind the fine-tuned GPT-3.5. Similarly, Alghisi et al. (2024) demonstrated that the effectiveness of adaptation techniques such as fine-tuning and retrieval-augmented generation (RAG) depends not only on the technique itself but also on the choice of base model. These findings emphasize the importance of careful model selection to achieve optimal fine-tuning outcomes.

Another limitation of current research is the insufficient multilingual understanding of domain-specific applications of LLMs. Most studies rely on English-language datasets, providing little insight into LLM performance in underrepresented languages. State-of-the-art large language models available today have primarily been developed in English due to the dominance of the English language in their training datasets. With approximately 6,500 spoken languages

worldwide (Hammarström, 2016), many languages are significantly underrepresented in LLM development due to the large data requirements (Muller et al., 2021). Several studies have investigated LLM performance across languages and consistently found a strong bias toward high-resource languages. Bang et al. (2023) also reported that ChatGPT struggles to generate sentences in low and extremely low-resource languages and can produce hallucinations. Lai et al. (2023) found that ChatGPT performs better overall in English than in the other 37 languages being examined. Ahuja et al. (2023) reported that LLMs perform better in high-resource languages and languages with Latin scripts in a study that involved 33 languages.

Prior studies also highlight that LLMs struggle with low-resource languages, particularly in capturing local linguistic and cultural nuances. For instance, Nazi et al. (2025) found that the open-source BLOOM model achieved only 11 % accuracy on Bengali text classification, reflecting significant difficulty in handling the language, while closed-source models (GPT-3.5, GPT-4o, Gemini) consistently outperformed open-source models (Aya 101, BLOOM, LLaMA) on reasoning-heavy tasks. Similarly, Buscaldi and Rosso (2023) reported that the NLLB-200 translation model struggles with the Genoese language, frequently misrepresenting local toponyms and culturally specific terms. Developed by Meta AI under the No Language Left Behind initiative, the NLLB-200 model aims to address gaps in Machine Translation for low- and very low-resource languages, but this case illustrates persistent challenges for underrepresented languages.

The Dutch datasets, although the precise language distribution in LLM training datasets is not publicly disclosed, can be estimated because the training data mostly comes from publicly available internet sources. Lin et al. (2022) reported that GPT-3 contains only 7% non-English text. In addition, Lai et al. (2023) estimated that, based on the latest Common Crawl dataset (Common Crawl Foundation, n.d.), only 2% of the data used for training GPT-3 is in Dutch. Within this portion, the overall Dutch data spans all topics and domains; thus, the amount of data specifically related to the transportation domain is likely much smaller. Consequently, although Dutch is not generally considered a low-resource language and is spoken by approximately 24 million people according to the Dutch Language Union (Taal Unie, n.d.), it can become effectively underrepresented in domain-specific contexts such as transportation.

In theory, generative language models can be trained for any language. In practice, however, developing such models for specific languages and domains is challenging. Vanroy (2023) noted that progress in generative language modeling for Dutch has been slow due to limited data, computational resources, and expertise in this rapidly evolving field. As a result, adapting existing LLMs remains a more practical and economical approach. A recent study by De Langhe et al. (2024) evaluated the performance of several large language models on Dutch general-purpose NLP tasks. The authors examined zero-shot learning across six tasks: sentiment analysis, emotion detection, irony detection, topic classification, and event coreference resolution. A range of models was evaluated, including natural language inference models, masked language models, and generative models (e.g., BERTje, mDeBERTaV3, robbert-v2-dutch-base, GPT-3.5 Turbo, and Llama 2…). These model performances were compared with the best supervised state-of-the-art (SOTA) Dutch models for each dataset as a benchmark. Among the evaluated models, GPT-3.5 Turbo achieved the best overall performance; however, it still lagged behind supervised SOTA systems in a zero-shot setting but was ahead in emotion detection. The study highlights the potential of large generative models but is limited to general text classification and text-only input, without addressing domain-specific or multimodal tasks. Moreover, recent frontier systems, such as GPT-4 Turbo, GPT-4o, Gemini, and DeepSeek, with stronger reasoning and multimodal capabilities, were not included.

Consequently, it remains unclear how modern frontier models perform on domain-specific Dutch tasks involving multimodal inputs and standardized downstream pipelines, particularly in operational domains such as transportation. This limitation is especially important in transportation systems, where traffic regulations, infrastructure conditions, and safety information are often communicated in local languages and context-specific terminology. Understanding how

LLMs perform under such domain-specific and linguistically underrepresented conditions is, therefore, an important step toward enabling practical applications of LLMs in transportation systems.

To address this gap, this study investigates the capabilities of frontier LLMs in the transportation domain, focusing on cross-lingual performance (Dutch vs. English), complex reasoning tasks, and multimodal understanding involving traffic signs and local regulations.

3 Method

Transportation usually involves not only textual information but also numerous visual elements to describe road infrastructure, signal signs, and warning boards. Various input types (text, image, video, audio) are convertible; for example, audio can be transcribed into text, videos can be segmented into image frames, and text can be converted into images like screenshots. Based on the ability to convert between formats and the variety of ways text and images can represent information, we focused our analysis on textual and image inputs. Gemini and GPT models were selected for their capability to handle multimodal data, enabling the processing of both text and image data, while DeepSeek models were employed exclusively for text-based datasets. To assess the models more thoroughly, we designed datasets that contained text and combined text and images with gradually increasing complexity. This setup aimed to allow us to explore how well the models could reason and make sense of more challenging input.

3.1 Experimental Setup

The aim is to understand the performance of LLMs on transportation topics in Dutch and across levels of reasoning complexity. It also seeks to assess LLM capabilities on both textual and text–image inputs. Accordingly, the datasets were developed to support these objectives.

Datasets

The datasets used in this study were originally derived from our teaching and training material, and the broad set of current courses offered at the local university. These questions are available mainly in Dutch (the local language) and some in English. They are all multiple-choice, with predefined solutions as the reference for assessing the models' performance. For courses available in both Dutch and English, the questions exist in pairs, with the English versions translated from Dutch. These translations were carefully reviewed by the course instructors, all of whom are native Dutch speakers with certified English teaching qualifications, ensuring accuracy and consistency. The primary target audience for these training courses includes both professional and personal drivers, as well as vulnerable road users such as children who walk or bike to school. The questions cover various topics, including road safety, road user behaviors, rules and regulations, driver fitness, and eco-driving.

Textual datasets

For text-based assessments, we used two types of datasets: BoolQ and MulAns. The sample questions are illustrated in Table 1. Both contain multiple-choice questions, but BoolQ includes only one correct answer, either 'true' or 'false'. In contrast, MulAns allows multiple correct answers, requiring more complex reasoning and domain knowledge, as the model must evaluate all options and identify the best combination. To evaluate the models' understanding of the Dutch compared to the English language, we divided the BoolQ dataset into two language categories: English (BoolQ_en) and Dutch (BoolQ_nl). The BoolQ_en dataset contains 258 questions, while the BoolQ_nl dataset includes 225 questions. To further explore the capabilities of LLMs in more complex contexts, we compare their performance in Dutch using the BoolQ_nl and MulAns datasets, both of which are in Dutch. BoolQ_nl presents a relatively simple task, with binary (True/False) questions and only one correct answer, allowing for a 50% chance of guessing

correctly. In contrast, MulAns includes questions with 3 to 12 answer options, substantially reducing the likelihood of correct random guessing and increasing task complexity.

Text and image datasets

Regarding the multimodal datasets, each question consists of text and an image that describes a situation or context. We assessed basic driving knowledge using the BasicKnowl dataset, which featured questions derived from Belgian driving license exams. These questions are paired with images of standard traffic signs, which can be sourced from the internet, such as from Wikipedia. An example of the questions in the BasicKnowl dataset is illustrated in Table 1.

Since such images are likely to be part of the training data for LLMs, we introduced additional complexity by including questions paired with our own images. Those questions with original images were created by our institute to promote traffic safety education for schoolchildren via a gamified e-learning platform named Route2School (R2S). These images, captured from real-world road contexts across Flanders (Belgium), are not expected to appear in any LLM training data, thus allowing for a more reliable assessment of the model's capacity to interpret unseen visual information. The questions are designed for primary and secondary school pupils, so the language is intentionally kept simple. Given the target audience of young pupils, the questions focus on scenarios involving pedestrians or cyclists. Two datasets were derived from these questions: KidsKnowl and KidsRiskMgmt. Both datasets contain images that complement the text, depicting traffic signals, signs, road situations, and infrastructure. However, the KidsRiskMgmt dataset requires more complex reasoning. It involves understanding traffic rules and regulations and applying them to specific situations to manage risk and ensure traffic safety. Examples of the questions in the KidsKnowl and KidsRiskMgmt datasets are illustrated in Table 1. In the text-image datasets, all questions are multiple-choice with only one correct option.

Table 1. Sample questions in each dataset

Dataset	Sample Question	Translation
BoolQ_en	True or False: The implementation of roundabouts in intersection redesign efforts universally reduces the speed of all types of vehicles. **Option 1: True** Option 2: False	
BoolQ_nl	Door de motorrem te gebruiken wordt de snelheid soepeler verlaagd en spaar je de remvoeringen. **Optie 1: Waar** Optie 2: Onwaar	By using the engine brake, the speed is reduced more smoothly, and you save the brake pads. **Option 1: True** Option 2: False
MulAns	Welk type weggebruikers classificeren als zwakke weggebruikers? **Optie 1: Voetgangers** **Optie 2: Fietsers** Optie 3: Auto's **Optie 4: Motorfietsen** Optie 5: Alle bovenstaande antwoorden	What types of road users are classified as "vulnerable road users"? **Option 1: Pedestrians.** **Option 2: Cyclists** Option 3: Cars **Option 4: Motorcyclists** Option 5: All of the above
BasicKnowl	What is the rule for turning left at this location? Option 1: All vehicles must turn left. **Option 2: Only cyclists can turn left.** Option 3: No one can turn left.	

		
KidsKnowl	Waar moet Charlie fietsen? **Optie 1: Aan de linkerkant van het pad** Optie 2: Aan de rechterkant van het pad Optie 3: Charlie mag kiezen 	Where should Charlie cycle? **Option 1: On the left side of the path.** Option 2: On the right side of the path. Option 3: Charlie can choose. 
KidsRiskMgmt	Charlie komt deze rijbaankussens tegen, wat moet hij doen? Optie 1: Charlie moet hier tussen deze 2 kussens fietsen Optie 2: Charlie moet hier over het rechtse kussen fietsen **Optie 3: Charlie moet hier rechts van het rechtse kussen fietsen** 	Charlie encounters these speed cushions, what should he do? Option 1: Charlie should cycle between these 2 cushions. Option 2: Charlie should cycle over the right cushion. **Option 3: Charlie should cycle to the right of the right cushion.** 

* The correct option is marked in bold

LLMs

We considered three prevalent foundation LLM families (i.e., not fine-tuned) for evaluation, including closed-source models, including Gemini (Gemini Team, 2023), GPT (OpenAI, 2024b), and open-source DeepSeek (DeepSeek, 2025). Due to cost and practical constraints, models were selectively used for each testing case. For the evaluation of text-based datasets, we employed nine models: three from the Gemini family: Gemini 1.0 Pro, Gemini-1.5-Flash, and Gemini-1.5-Pro; four from the GPT family: GPT-3.5-Turbo, GPT-4o, GPT-4-Turbo, and GPT-o1 (used only for the MulAns dataset due to its considerably higher complexity); and two from the DeepSeek family: DeepSeek-V3 and DeepSeek-R1. For multimodal dataset evaluation, we used Gemini-1.5-Flash, Gemini-1.5-Pro, and GPT-4o, as some models do not support image processing and practical constraints. Each of these models was accessed using its corresponding API key. Specifically, we queried Gemini through Google Vertex AI and the GPT models through the OpenAI API and DeepSeek via the Deepinfra API platform (Deep Infra, 2025). We set the LLM models to be highly deterministic for all evaluations by using a temperature of 0. Additionally, we set the system role to “You are a transportation expert. Your task is to answer the given multiple-choice questions" during model response generation.

Prompt settings

We employed two main prompting settings: zero-shot standard prompting (0-shot SP) (Kojima et al., 2022) and instructional prompting (IP). The zero-shot prompt setting provides no additional text or context and aims to evaluate the models' inherent capabilities. In contrast, the IP setting adds a prefix to the prompt before questions, providing context and specifying the expected answer format. This approach is primarily intended to guide the model toward producing consistent responses in the correct format, which is necessary for automation and application development. Instruction prompting combined with a few-shot example setting was also evaluated on the text-based datasets to examine whether it could further improve model performance. The IP_en and IP_nl are the instructions in English and Dutch, only test with BoolQ questions in English and Dutch to better understand the self- and cross language command instructions.

Table 2. Overview of the datasets and the experimental setup

		Datasets					
		BoolQ_en	**BoolQ_nl**	**MulAns**	**BasicKnowl**	**KidsKnowl**	**KidsRiskMgmt**
Data	Nr of questions	258	225	183	50	130	126
	Language	English	Dutch	Dutch	English	Dutch	Dutch
	Type	text	text	text	text & image	text & image	text & image
LLM models	Gemini-1.0-pro	x	x	x			
	Gemini-1.5-flash	x	x	x	x	x	x
	Gemini-1.5-pro	x	x	x	x	x	x
	GPT-3.5-turbo	x	x	x			
	GPT-4-turbo	x	x	x			
	GPT-4o	x	x	x	x	x	x
	GPT-o1			x			
	DeepSeek V3	x	x	x			
	DeepSeek R1	x	x	x			
Prompt settings	0-shot	x	x	x	x	x	x
	IP_en	x	x	x	x	x	x
	IP_nl	x	x				
	k-shot (k=6) & IP_en	x	x	x			

Table 2 presents the experimental setup, showing the selected combinations of datasets, models, and prompt configurations used in this study. The design is guided by the main research objectives, including cross-lingual performance evaluation, complex reasoning assessment, and multimodal understanding of traffic signs and local regulations. Due to practical computational constraints, not all possible model–dataset–prompt combinations were tested. Instead, we selected targeted configurations that are most relevant to the study goals, allowing meaningful and focused comparisons.

Given that road signs and government road safety guidelines are often in the local language, assessing how well LLMs perform in local languages is crucial. To this end, we evaluated language performance in both Dutch and English using the BoolQ datasets. The instructional prompts were

provided in either English or Dutch. Data were tested with both self-lingual and cross-lingual instructional prompting to compare language performance. In self-lingual instructional prompting, the IP's language matched the question's language: English content questions in the BoolQ_en dataset were tested with the IP in English, and Dutch content questions in the BoolQ_nl dataset were tested with the IP in Dutch. In cross-lingual instructional prompting, the IP's language differed from the question's language: English content questions were tested with the IP in Dutch, and Dutch content questions were tested with the IP in English. The results of cross-lingual instructional prompting were compared to those of self-lingual instructional prompting.

Additionally, a few-shot prompting with six examples combined with IP was implemented for the BoolQ and MulAns datasets to explore potential enhancements in model performance. For the multimodal dataset, we tested with zero-shot prompting and IP in English to observe how performance depended on the prompting settings.

3.2 Evaluation method

Since all the questions are multiple-choice questions, with a clear ground truth, the evaluation process is straightforward and largely objective. For the 0-shot SP evaluation, we employed human assessment, where the evaluator recorded the option selected by the LLM for each question. Three evaluators participated: one author and two assistants, a student and a staff member from the institute. To evaluate the settings with IP and IP-few-shot, we evaluated the responses automatically using Python scripts since the responses must follow a standard format.

The responses given by the LLM could be evaluated in two ways: *global accuracy* and *local accuracy*. The global accuracy evaluation compares the response of the LLM to the benchmark response to see if they match exactly. When they match exactly, the response of the LLM is labeled "correct".

The global accuracy is measured by Equation 1:

$$Global\ Accuracy\ (\%) = \frac{total\ correct\ responses}{total\ questions} * 100 \quad (1)$$

To assess performance at a more granular level, local accuracy evaluates each option within a question individually, rather than requiring all selected options to be correct, as in global accuracy. It measures the proportion of matched options, even if the entire answer is not fully accurate. For this purpose, the LLM's response is transformed into a binary index vector, where each position corresponds to an option: a value of '1' indicates that the option was selected, and '0' that it was not. This LLM's response vector is then compared to the ground truth vector to determine the number of correct matches. The local accuracy is computed as shown in Equation 2. Since global accuracy only considers completely correct answers, it is always less than or equal to local accuracy. In Boolean questions, however, global and local accuracy are equal due to the mutually exclusive nature of the choices (i.e., either true or false).

$$Local\ Accuracy\ (\%) = \frac{\sum nr\ of\ correctly\ matched\ indices\ per\ response}{\sum nr\ of\ indices\ per\ question} * 100 \quad (2)$$

Table 3 illustrates how global and local accuracy are calculated using a sample question: *'What is allowed when under the influence of alcohol as a cyclist?'*. The question provides three answer options.

The correct answer includes Option 1 and Option 3, which can be represented as an index vector: [1, 0, 1]. The LLM's response selects only Option 3, corresponding to the vector [0, 0, 1].

Global accuracy requires all selected options to match the benchmark exactly. Since the LLM's response omits Option 1, the answer is considered incorrect, resulting in a global accuracy of 0% for this question. In contrast, local accuracy evaluates each option independently. In this example, the LLM correctly selected Option 3 and correctly did not select Option 2 but failed to select Option 1. Out of three options, two were correctly matched, giving a local accuracy of 67% (2 out of 3). In this study, global accuracy was used to evaluate BoolQ and the text-image datasets (BasicKnowl,

KidsKnowl, and KidsRiskMgmt), while both global and local accuracy evaluations were applied to assess the MulAns dataset.

Table 3. Example of global and local accuracy

	Question	Correct Answer Benchmark	Response from an LLM	Global Accuracy	Local Accuracy
	What is allowed when under the influence of alcohol as a cyclist? **Option 1: Leave the bicycle behind and walk home.** Option 2: Proceed to cycle home. **Option 3: Push the bicycle home.**	**Option 1, Option 3**	"The correct answer is **Option 3**: Push the bicycle home. Riding a bicycle under the influence of alcohol is dangerous and often illegal, similar to driving a motor vehicle while intoxicated. Pushing the bicycle home while walking is a safer and typically legal alternative, as it avoids the risks associated with impaired balance, judgment, and motor skills that come with cycling under the influence. Option 1, leaving the bicycle behind, might not be safe for the bicycle, and Option 2, proceeding to cycle home, is unsafe and likely unlawful."	0%	67%
Index vector format		[1,0,1]	[0,0,1]	The full set doesn't match.	Two out of three items matched

* The correct option is marked in bold

Alongside the local accuracy, the F1, precision, and recall were employed to provide a more detailed assessment of the LLM's errors in binary true/false classification. The metrics are computed by aggregating the total true positives (TP), false positives (FP), and false negatives (FN) across all questions:

$$\mathrm{TP} = \sum_{q} \mathrm{TP}_q$$

$$\mathrm{FP} = \sum_{q} \mathrm{FP}_q$$

$$FN = \sum_{q} \mathrm{FN}_q$$

Then the precision, recall, and F1 scores are calculated as follows:

$$\text{Precision} = \frac{TP}{TP + \mathrm{FP}}$$

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

There were a few instances where the models declined to respond to certain queries, particularly those flags potentially illegal or unethical content. When a response triggers an error flag, the system attempts the query up to three times. Occasionally, models provide answers that fall outside the expected options or include conflicting information within a single response. These cases are categorized as unanswered or incorrect.

For answers that do not follow the required IP format, we categorize them as incorrect. All questions were written as fully self-contained questions, meaning they can be understood without additional instructions. The instructions included in the test case are designed to evaluate the ability of LLMs to produce answers in a consistent format, which is beneficial for automated pipelines. Therefore, responses with an incorrect format are categorized as incorrect answers. In such cases for global accuracy, the number of correct indices for that answer is set to 0, while the total number of indices in the question is still counted.

For the settings in GPT and DeepSeek, such invalid answers were very rare, often 0%, with some around 1-2%. The Gemini model shows a slightly higher rate of flagged responses, primarily due to safety checks, but this is mostly limited to one or two specific prompt settings in MulAns database, with the maximum occurrence of 10%.

4 Results

4.1 Language impact on LLM's performance

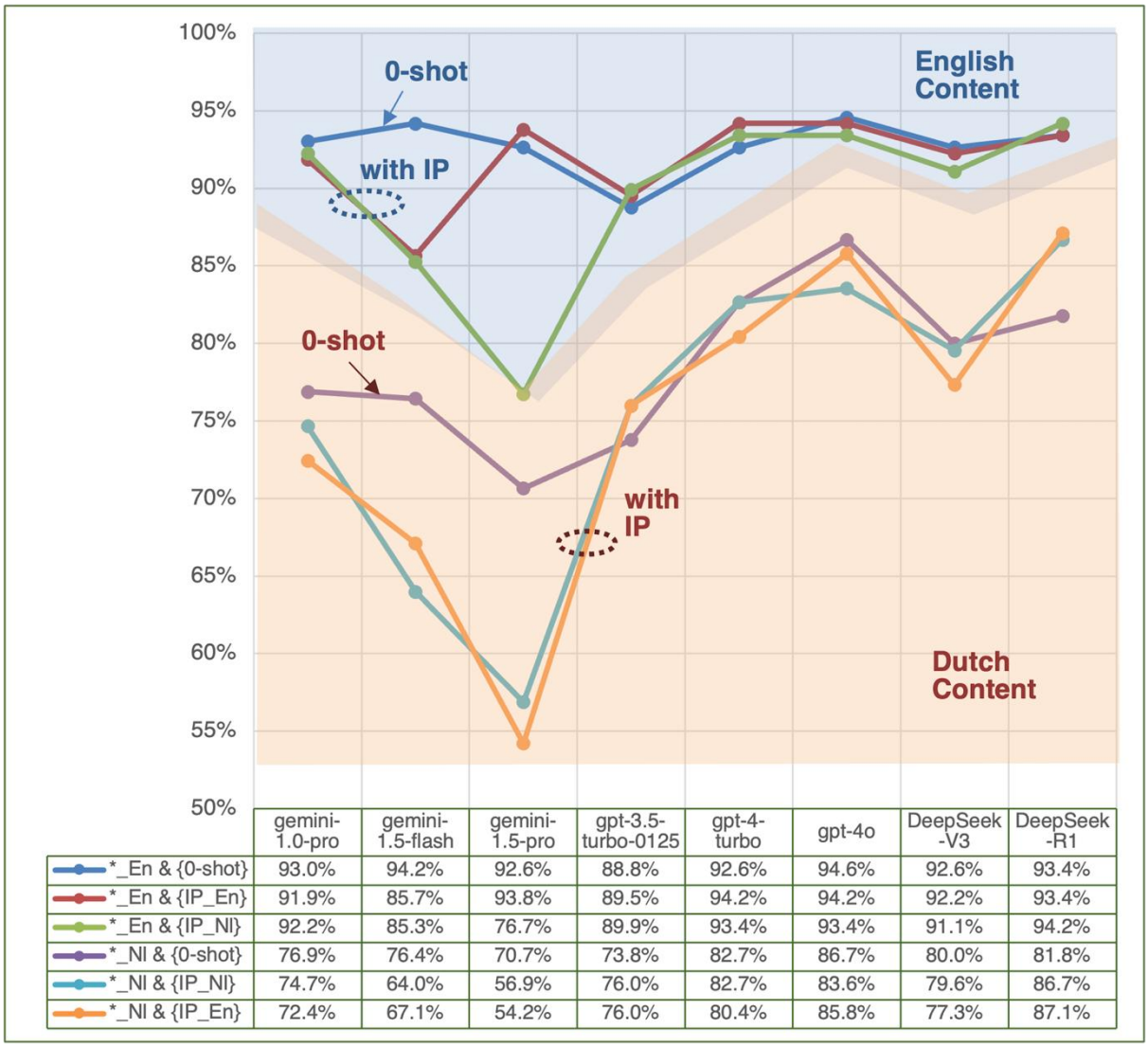
Figure 1 illustrates the performance of LLMs on two Boolean question-answering datasets in different languages: BoolQ_en (English) and BoolQ_Nl (Dutch). The models were evaluated under three prompting conditions: (i) zero-shot, (ii) self-lingual instruction-based prompting, where the language of the instructional prompt (IP) matched the input language (i.e., English content with an English prompt (*_En & IP_En), Dutch content with a Dutch prompt (*_Nl & IP_Nl)), and (iii) cross-lingual instruction-based prompting, where the instructional prompt was in the different language of the input (i.e., English content with a Dutch prompt (*_En & IP_Nl), and vice versa (*_Nl & IP_En)). The cross-lingual condition aimed to test model robustness when the prompt language differed from the task language. This scenario could arise in real-world applications, where English prompts were often used regardless of input languages. The evaluation included eight LLMs from three model families: Gemini (gemini-1.0-pro, gemini-1.5-flash, gemini-1.5-pro), GPT (gpt-3.5-turbo, gpt-4-turbo, gpt-4o), and DeepSeek (DeepSeek-V3, DeepSeek-R1). The performance was assessed based on the global accuracy as illustrated in Equation 1. Together with Figure 1, Table 4 provides complementary information about the performance gaps of LLMs between English and Dutch. It compares LLM performance in a zero-shot setting on the English dataset (BoolQ_en) and the Dutch dataset (BoolQ_nl), then calculates the average performance for each model family.

There are several key takeaways. From a language perspective, LLMs generally performed better on English content than on Dutch content across all models and prompt settings, as summarized in two different color zones in Figure 1. With zero-shot prompting, LLM performance on the English dataset BoolQ_en ranged from 88.8% (GPT-3.5 Turbo) to 94.6% (GPT-4o). In contrast, on the Dutch dataset BoolQ_nl, performance ranged from 70.7% (Gemini 1.5 Pro) to 86.7% (GPT-4o).

Table 4 provides a detailed view of performance gaps across different languages. In the 0-shot setting, where the base models were evaluated without additional prompting, the performance difference between English and Dutch content exceeded 10 percentage points for all tested LLMs, except GPT-4o, which showed the smallest gap at 7.9 percentage points. In contrast, Gemini 1.5 Pro exhibited the most significant gap at 21.9 percentage points. On average, the GPT family performed best across language content, showing an average difference of 11 percentage points between English and Dutch content. DeepSeek followed closely with a 12.1 percentage point difference, while the Gemini models showed the largest gap, averaging 18.6 percentage points.

The added value of instruction prompting (IP), whether applied in the same language (self-lingual) or in another language (cross-lingual), was generally comparable. Across both datasets, BoolQ_en and BoolQ_nl, the performance patterns of the two approaches were closely aligned, as shown in Figure 1. The only clear deviation was observed for Gemini 1.5-Pro. On the English BoolQ_en

dataset, this model exhibited a marked sensitivity to the language of instruction: accuracy achieved 93.8% with English IP, but dropped sharply to 76.7% with Dutch IP. By contrast, on the Dutch BoolQ_nl dataset, its performance remained relatively stable across instruction languages, reaching 54.2% with English IP and 56.9% with Dutch IP.



	gemini-1.0-pro	gemini-1.5-flash	gemini-1.5-pro	gpt-3.5-turbo-0125	gpt-4-turbo	gpt-4o	DeepSeek-V3	DeepSeek-R1
*_En & {0-shot}	93.0%	94.2%	92.6%	88.8%	92.6%	94.6%	92.6%	93.4%
*_En & {IP_En}	91.9%	85.7%	93.8%	89.5%	94.2%	94.2%	92.2%	93.4%
*_En & {IP_Nl}	92.2%	85.3%	76.7%	89.9%	93.4%	93.4%	91.1%	94.2%
*_Nl & {0-shot}	76.9%	76.4%	70.7%	73.8%	82.7%	86.7%	80.0%	81.8%
*_Nl & {IP_Nl}	74.7%	64.0%	56.9%	76.0%	82.7%	83.6%	79.6%	86.7%
*_Nl & {IP_En}	72.4%	67.1%	54.2%	76.0%	80.4%	85.8%	77.3%	87.1%

Figure 1. Performance comparison of eight LLMs across two different languages in content and instructional prompts

Table 4. Performance gap between English and Dutch languages in different settings

Models	0-shot Perf. Gap (En vs. Nl content)	Avg. Perf. Gap per LLM Family
gemini-1.0-pro	16.1%	18.6%
gemini-1.5-flash	17.7%	
gemini-1.5-pro	21.9%	
gpt-3.5-turbo	15.0%	11.0%
gpt-4-turbo	10.0%	
gpt-4o	7.9%	
DeepSeek-V3	12.6%	12.1%
DeepSeek-R1	11.6%	

The impact of IP compared to 0-shot prompting, however, varied depending on the model and the language of the dataset. For the GPT models, self-lingual and cross-lingual instructional prompting generally led to better performance than 0-shot prompting for GPTs, except for GPT-4o, where its performance declined in both datasets, or for GPT-4 Turbo when it performed in the BoolQ_nl dataset in the cross-lingual setting. In contrast, the three Gemini models typically experienced a performance decline with instructional prompting compared to the zero-shot setting. Gemini 1.0 Pro showed a slight decline across all the settings, while Gemini 1.5 Flash displayed more pronounced decreases, with the most significant drop of 12.4 percentage points on BoolQ_nl and Dutch IP (76.4% vs 64.0%). Gemini 1.5 Pro showed mixed results, with a slight improvement on BoolQ_en using English IP but significant declines elsewhere, with the largest decline of 16.5 percentage points on BoolQ_nl with English IP (70.7% vs. 54.2%). DeepSeek-V3 did not benefit from IP, exhibiting a slight reduction in performance compared to zero-shot prompting. On the other hand, DeepSeek-R1 showed notable improvements with IP, particularly for Dutch content (or the BoolQ_nl dataset). Instructional prompting enhanced performance from 81.8 % accuracy in zero-shot prompting to 86.7% in self-lingual IP (Dutch prompts) and to 87.1% in cross-lingual IP (English prompts).

On average, GPT and DeepSeek models handled different content languages better than Gemini models. Among the models tested, DeepSeek-R1 and GPT-4o exhibited the strongest multilingual capabilities in handling Dutch and English under zero-shot prompting. With instructional prompting in both self- and cross-lingual settings, DeepSeek-R1 demonstrated additional performance gains.

4.2 Overall model performance on text-based inputs

Figure 2 presents the global accuracy (GA) comparison among several large language models on two text-based Dutch-language evaluation datasets: BoolQ_nl (Boolean question answering) and MulAns (multiple-choice question answering). The models were assessed under three prompting conditions: (i) zero-shot (0-shot), (ii) instruction-based prompting in English (IP), and (iii) instruction and few-shot prompting in English with six exemplars in both English and Dutch (IP & 6-shot). The evaluation included eight LLMs from three model families: Gemini (gemini-1.0-pro, gemini-1.5-flash, gemini-1.5-pro), GPT (gpt-3.5-turbo, gpt-4-turbo, gpt-4o), and DeepSeek (DeepSeek-V3, DeepSeek-R1). For the MulAns dataset, an additional evaluation was conducted using GPT-o1, a reasoning-optimized variant, to examine its comparative performance on more complex tasks relative to DeepSeek -R1.

The results reveal several key trends. All evaluated LLMs outperformed random guessing for the Boolean reasoning tasks, where each question had a binary (true/false) answer, corresponding to the chance-level performance of 50%. GPT-4o, GPT-4 Turbo, and both DeepSeek models consistently achieved the highest accuracy across all prompt settings and input languages, as shown in Figure 2.

In contrast, performance on the multiple-choice question (MulAns) dataset was considerably lower, reflecting the increased complexity of this task. The MulAns format involved up to 12 answer options per question, significantly reducing the likelihood of correct answers by chance. Under the zero-shot prompt, most models performed poorly, with global accuracy falling below 33%. An exception was GPT-o1, which achieved a better accuracy of 44.3%. The weakest performance was observed with GPT-3.5 Turbo, which answered only 11.5% of the questions correctly.

There was a clear performance progression across successive versions of the GPT and DeepSeek models, observed via the 0-shot in two datasets, with accuracy consistently improving from GPT-3.5 Turbo to GPT-4o and further to GPT-o1 on the MulAns dataset. DeepSeek R1 also performed better than DeepSeek V3 on both datasets. In contrast, the Gemini models exhibited inconsistent performance trends across versions. For instance, in the BoolQ_nl dataset under the zero-shot

condition, the oldest model, Gemini 1.0 Pro, achieved the highest accuracy (72.4%), outperforming both Gemini 1.5 Flash (67.1%) and Gemini 1.5 Pro (54.2%).

Regarding instructional prompting, while adding instructional prompts improved the readability and structure of responses, it did not yield a consistent performance gain on the BoolQ_nl dataset with GPT and DeepSeek models. However, across all Gemini models, IP alone resulted in lower performance than zero-shot prompting, but performance improved substantially when examples were included. On the more complex dataset MulAns, instructional prompting enhanced performance across all models, as shown in Figure 2. However, the marginal difference between using an instructional prompt alone and combining it with k-shot examples (k = 6) was not consistently significant.

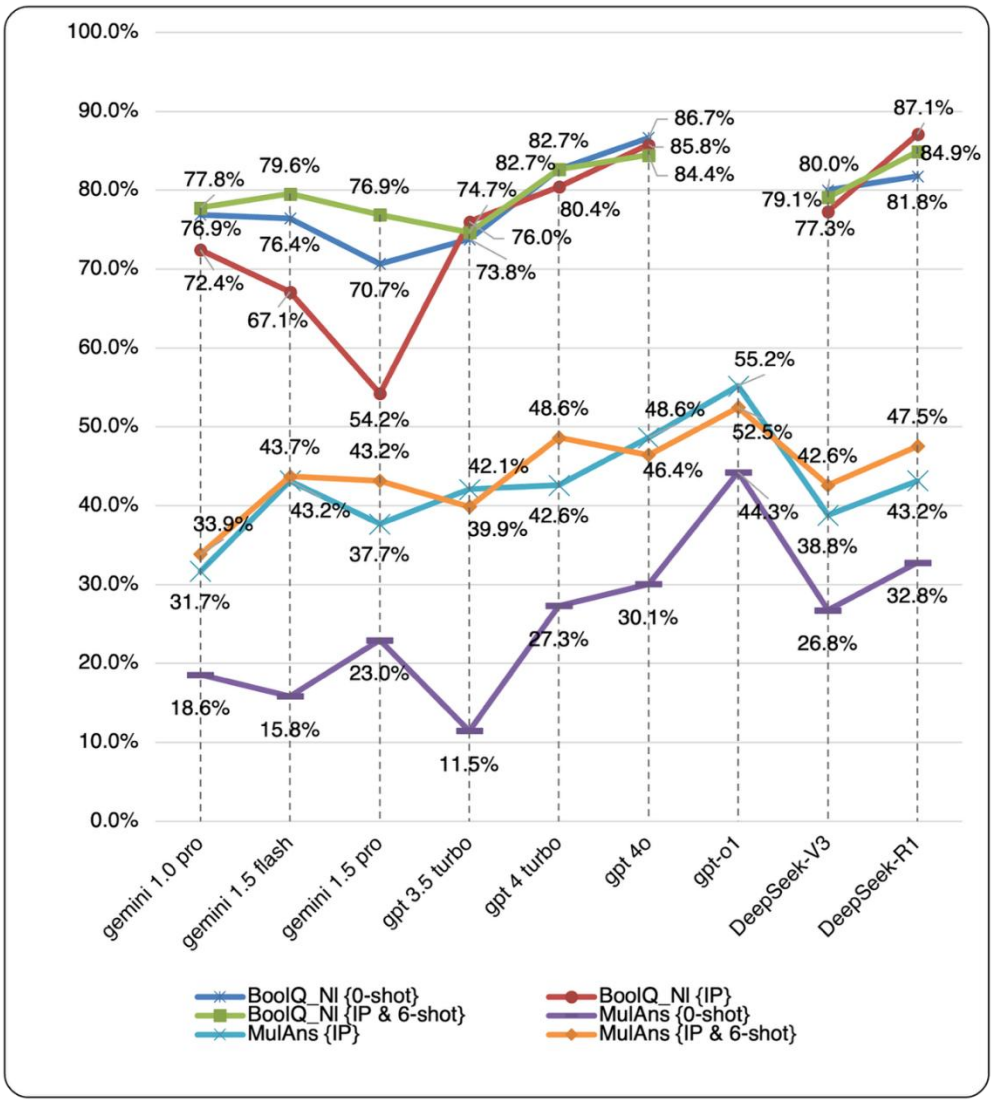
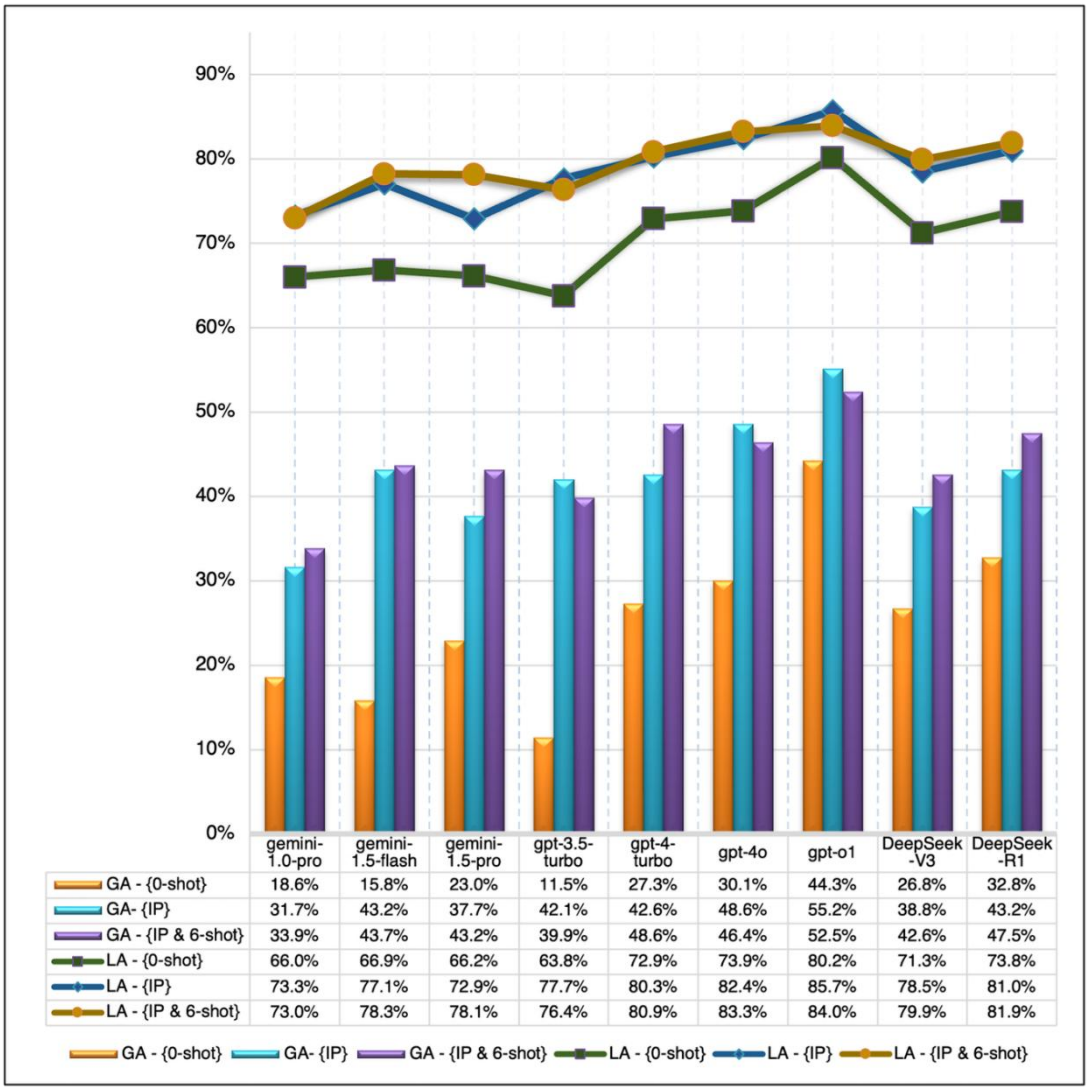


Figure 2. LLM's Performance comparison across different text datasets

To gain better insight into the MulAns dataset's response accuracy, global and local accuracy were calculated using Equations 1 and 2, respectively, which are presented in Figure 3. Global accuracy provides an overall performance measure but often appears low due to the difficulty of multiple-choice questions, with multiple correct options, which could form numerous possible combinations. On the other hand, local accuracy provides a complementary insight into how far the LLM is progressing in providing a correct answer, giving a more nuanced view of its performance.

Looking at local accuracy trajectories in Figure 3, all models achieved local accuracies above 50%, which is the baseline probability of random guessing when an option is selected randomly. Additionally, local accuracy generally aligned with global accuracy across all models and prompt settings. However, the gap between global and local accuracy revealed underlying progress. For instance, GPT-3.5 Turbo demonstrated significantly lower global accuracy than Gemini 1.5 Pro (11.5% vs. 23%) but had similar local accuracies (63.8% vs. 66.2%). This discrepancy indicated a potential for improvement. For instance, GPT-3.5 Turbo global performance improved with instructional prompting, reaching 42.1%, which surpassed Gemini 1.5 Pro's 37.7%. GPT-o1 consistently outperformed all other models across the MulAns dataset and prompt settings, achieving the highest global accuracy (GA) of 55.2% and an impressive local accuracy of 85.0%. The followers were DeepSeek-R1 and GPT-4o, which attained peak global accuracies of 47.5% and 48.6%, respectively.



	gemini-1.0-pro	gemini-1.5-flash	gemini-1.5-pro	gpt-3.5-turbo	gpt-4-turbo	gpt-4o	gpt-o1	DeepSeek-V3	DeepSeek-R1
GA - {0-shot}	18.6%	15.8%	23.0%	11.5%	27.3%	30.1%	44.3%	26.8%	32.8%
GA- {IP}	31.7%	43.2%	37.7%	42.1%	42.6%	48.6%	55.2%	38.8%	43.2%
GA - {IP & 6-shot}	33.9%	43.7%	43.2%	39.9%	48.6%	46.4%	52.5%	42.6%	47.5%
LA - {0-shot}	66.0%	66.9%	66.2%	63.8%	72.9%	73.9%	80.2%	71.3%	73.8%
LA - {IP}	73.3%	77.1%	72.9%	77.7%	80.3%	82.4%	85.7%	78.5%	81.0%
LA - {IP & 6-shot}	73.0%	78.3%	78.1%	76.4%	80.9%	83.3%	84.0%	79.9%	81.9%

Figure 3. Global accuracy (GA) and local accuracy (LA) of the MulAns dataset

To better understand the errors made by the LLMs, precision, recall, and F1 were calculated using Equation 3, treating each multiple-choice option as a binary (true/false) decision. By aggregating all true positives, false positives, and false negatives before computation, these metrics offer a more detailed view of model performance.

Table 5 presents the F1, precision, and recall for all models under 0-shot, instruction prompting (IP), and IP with few-shot examples (k=6). Overall, the results indicated strong performance, with most models achieving strong scores across all three metrics. Among the tested models, GPT-o1

performed best overall across prompting strategies, consistently achieving the highest F1 scores. In the zero-shot setting, it reached an F1 score of 0.84 (precision = 0.84, recall = 0.84), increased to an F1 score of 0.89 (precision = 0.85, recall = 0.94) under instruction prompting, and remained strong with few-shot examples (F1 = 0.88, precision = 0.83, recall = 0.93). GPT-4o and DeepSeek-R1 followed closely, demonstrating consistent performance across prompting strategies.

Table 5. The F1, precision, and recall across different LLMs and prompting strategies

	0-shot			Instruction Prompting (IP)			IP & few-shot (k=6)		
Model name	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
gemini-1.0-pro	0.72	0.75	0.70	0.81	0.75	0.88	0.81	0.75	0.90
gemini-1.5-flash	0.70	0.81	0.62	0.82	0.80	0.84	0.83	0.82	0.85
gemini-1.5-pro	0.74	0.83	0.67	0.84	0.81	0.87	0.83	0.81	0.85
gpt-3.5-turbo	0.66	0.78	0.57	0.82	0.84	0.80	0.81	0.81	0.82
gpt-4-turbo	0.75	0.85	0.67	0.84	0.83	0.85	0.85	0.84	0.85
gpt-4o	0.77	0.83	0.72	0.86	0.83	0.91	0.87	0.83	0.92
gpt-o1	**0.84**	0.84	**0.84**	**0.89**	**0.85**	**0.94**	**0.88**	0.83	**0.93**
DeepSeek-V3	0.75	0.81	0.69	0.83	0.79	0.88	0.84	0.81	0.88
DeepSeek-R1	0.77	0.84	0.71	0.85	0.82	0.88	0.86	0.83	0.88

Instruction prompting improved F1 scores for all models, although the magnitude of improvement varied. The stronger models, which had already performed well in zero-shot settings, showed modest gains. For example, GPT-o1 improved from an F1 score of 0.84 to 0.89 (with a gain of +0.05), and GPT-4o from an F1 score of 0.77 to an F1 score of 0.86 (+0.09). By contrast, models with lower zero-shot performance benefited more substantially: GPT-3.5-turbo rose from an F1 score of 0.66 to 0.82 (+0.16), and Gemini-1.5-flash from an F1 score of 0.70 to 0.82 (+0.12). The effect of instruction prompting was especially pronounced in recall. Across models, average precision increased only slightly (average change approximately equals 0), whereas average recall rose markedly (average change is 0.17). For instance, GPT-3.5-turbo's recall improved from 0.57 to 0.80 (+0.23) with only a 0.06 increase in precision (0.78 to 0.84), and Gemini-1.0-pro's recall increased from 0.70 to 0.88 (+0.18) without any change in precision. Therefore, the instruction prompting improved the recall score, leading to better F1-scores overall. These results suggested that although instruction prompting consistently enhanced performance, its relative impact was greater for models starting with lower zero-shot baselines.

Shifting from instruction prompting to instruction plus few-shot prompting yielded only marginal changes in F1 scores, typically ±0.01. Some models experienced slight improvements (e.g., GPT-4o: 0.86 → 0.87), while others showed minor decreases (e.g., GPT-o1: 0.89 → 0.88). Overall, few-shot examples contributed minimally beyond the gains achieved through instruction prompting alone.

Furthermore, compared with the local accuracy in Figure 3, the precision aligned closely with the local accuracy of the models. For instance, in IP-k-shot, models like gpt-4o and gpt-o1 achieved local accuracies of 0.83 and 0.84, respectively, which closely matched their precision (both 0.83). At the same time, F1 and recall remained strong, showing that the models not only perform well in terms of accuracy but also reliably select correct options.

4.3 Performance on text and image input

Accessing transportation domain knowledge is more effectively achieved through the multimodal input dataset, as it combines text and images to better explain contextual factors such as road types,

roundabouts, traffic signs, and types of road users. The results of the multimodal input experiments are illustrated in Figure 4.

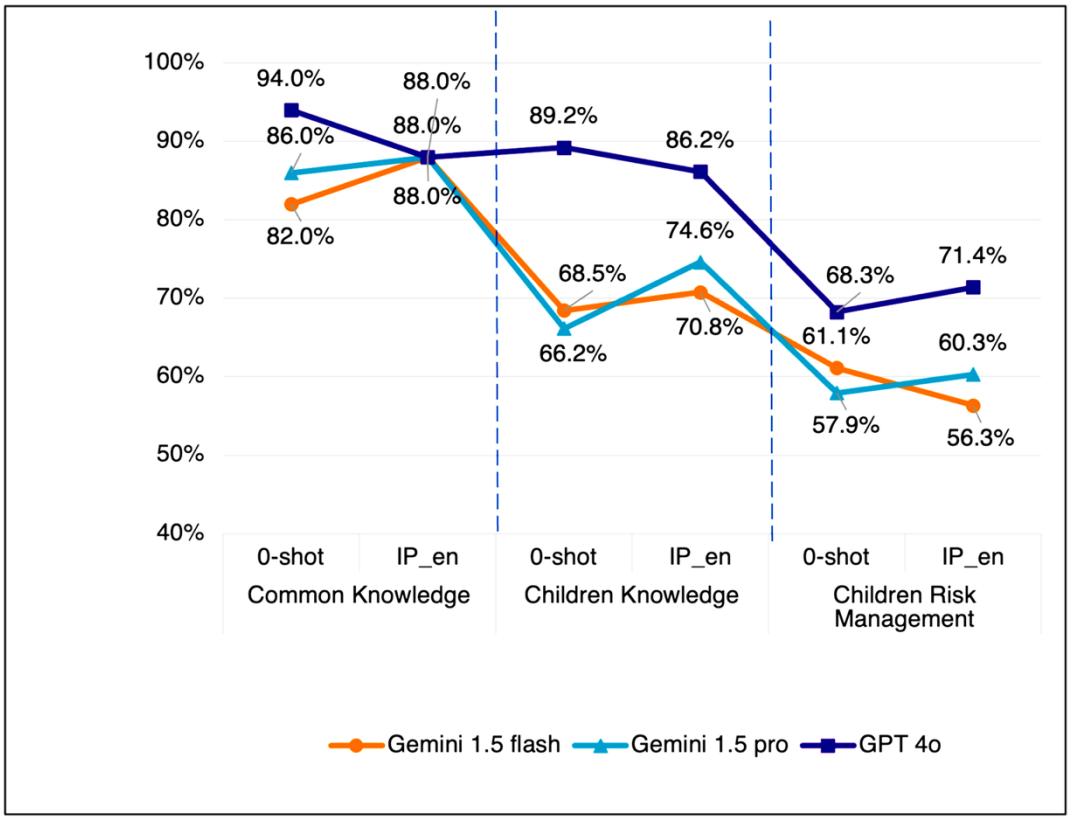


Figure 4. Performance comparison of three LLMs across different multimodal input datasets

The experiment results show that GPT-4o outperformed the Gemini models across all three datasets. The two Gemini models exhibited similar performance across different models and prompt settings. For basic transportation knowledge, where relevant images came from Wikipedia and may have appeared in the training data of the LLMs, the three models performed well in 0-shot SP, achieving accuracies of 82%, 86%, and 94% for Gemini 1.5 Flash, Gemini 1.5 Pro, and GPT-4o, respectively. When English instructions were added, the performance of the Gemini models improved to 88%, while GPT-4o's performance decreased. A similar trend was observed for the KidsKnowl dataset, which consisted of questions about basic knowledge at a children's level of understanding. The images used in this dataset came from private resources and were not part of the LLMs' training data, demonstrating the models' ability to apply knowledge in new contexts. In 0-shot SP, Gemini 1.5 Flash and Gemini 1.5 Pro achieved accuracies of 68.5% and 66.2%, respectively, substantially lower than the dataset with images from the internet. The performance increased to 70.8% and 74.6% with instructional prompting.

In contrast, GPT-4o's performance in 0-shot SP was 89.2%, which decreased to 86.2% with instructional prompting. In the previous basic knowledge datasets, Gemini 1.5 Pro achieved higher accuracy than Gemini 1.5 Flash in 0-shot SP. However, in the Risk Management dataset, Gemini 1.5 Flash performed better, with an accuracy of 61.1% compared to 57.9% for Gemini 1.5 Pro. Although designed for a children's understanding level, this dataset requires more advanced reasoning and decision-making, as it involves applying transportation knowledge to assess situations and actions. With the introduction of instructional prompting, Gemini 1.5 Pro showed improved results, reaching 60.3% accuracy versus 56.3% for Gemini 1.5 Flash. Despite this, GPT-4o demonstrated superior performance, with an accuracy of 68.3% in 0-shot SP and further improvement to 71.4% with instructional prompting.

5 Discussion

5.1 Beyond Random

LLMs demonstrate promising performance in the transportation domain even without additional fine-tuning. This is evident from their high response accuracy across various in-context transportation datasets, significantly outperforming random guessing, as shown by high *global accuracy* in BoolQ datasets and high *local accuracy* in MulAns, above 50%, compared to random guessing. The strong precision, recall, and F1 scores on the MulAns dataset reemphasized that the LLMs are not only achieving reasonable accuracy but also reliably identifying correct options.

A closer look at the reasoning behind their responses reveals that their accuracy is not solely based on simple reasoning within the text but is also informed by the extensive information they were trained on prior to deployment while answering questions, as illustrated in Figure 5.

<table>
<tr><td colspan="2">Question: “True or False: Electric and hybrid vehicles inherently reduce the risk of all types of crashes due to their sustainable design.
Option 1: True; Option 2: False”</td></tr>
<tr><td colspan="2">Response from GPT 4o:
“Option 2: False
Electric and hybrid vehicles are designed to be more environmentally friendly and energy-efficient, but this does not inherently reduce the risk of all types of crashes. The risk of crashes is influenced by a variety of factors, including driver behavior, road conditions, and vehicle safety features. While electric and hybrid vehicles may have advanced safety technologies, their sustainable design alone does not guarantee a reduction in crash risk."</td></tr>
<tr><td>Question:
“Welk type weggebruikers classificeren als zwakke weggebruikers?
Optie 1: voetgangers
Optie 2: fietsers
Optie 3: auto’s
Optie 4: motorfietsen
Optie 5: Alle bovenstaande antwoorden”</td><td>Translated:
“What types of road users are classified as “vulnerable road users”?
Option 1: Pedestrians
Option 2: Cyclists
Option 3: Cars
Option 4: Motorcyclists
Option 5: All of the above”</td></tr>
<tr><td colspan="2">Response from Gemini 1.5 pro:
“Answer(s): Option 1, Option 2, Option 4
Explanation: "Kwetsbare weggebruikers" translates to "vulnerable road users" in English. In Belgian traffic regulations, this term typically includes:
***Option 1: Voetgangers (Pedestrians):** They are the most vulnerable as they have no physical protection.
***Option 2: Fietsers (Cyclists):** While they have some protection from helmets, they are still vulnerable compared to motorized vehicles.
***Option 4: Motorrijders (Motorcyclists):** Although they operate a motorized vehicle, they lack the external protection of a car.”</td></tr>
</table>

Figure 5. Examples of LLMs' responses for the BoolQ_en and MulAns datasets show their reasoning capabilities. Note: the correct option is marked in bold.

5.2 Language Bias

The explanation of the response of LLMs also reveals their cross-lingual and multimodal generation capabilities, meaning they can apply knowledge from one language to another and interpret different types of inputs. In our case study, the model understood questions in Dutch and linked them to relevant knowledge. This capability was illustrated in an example from the BasicKnowl dataset in Figure 6, where the question was in English, but the image contained words in Dutch. GPT-4o understood the information on a signal board, accurately interpreting a turning

sign that meant "No left turn" and a bicycle sign. It comprehended the Dutch word "UITGEZONDERD," meaning "except," and used this information to reason and choose the correct solution. In a similar example, Gemini read and understood the text in Dutch and interpreted the information in the images about the board regarding cyclists and pedestrians, as well as the road layout from the photo. It connected this information to the question, enabling it to reason and answer correctly.

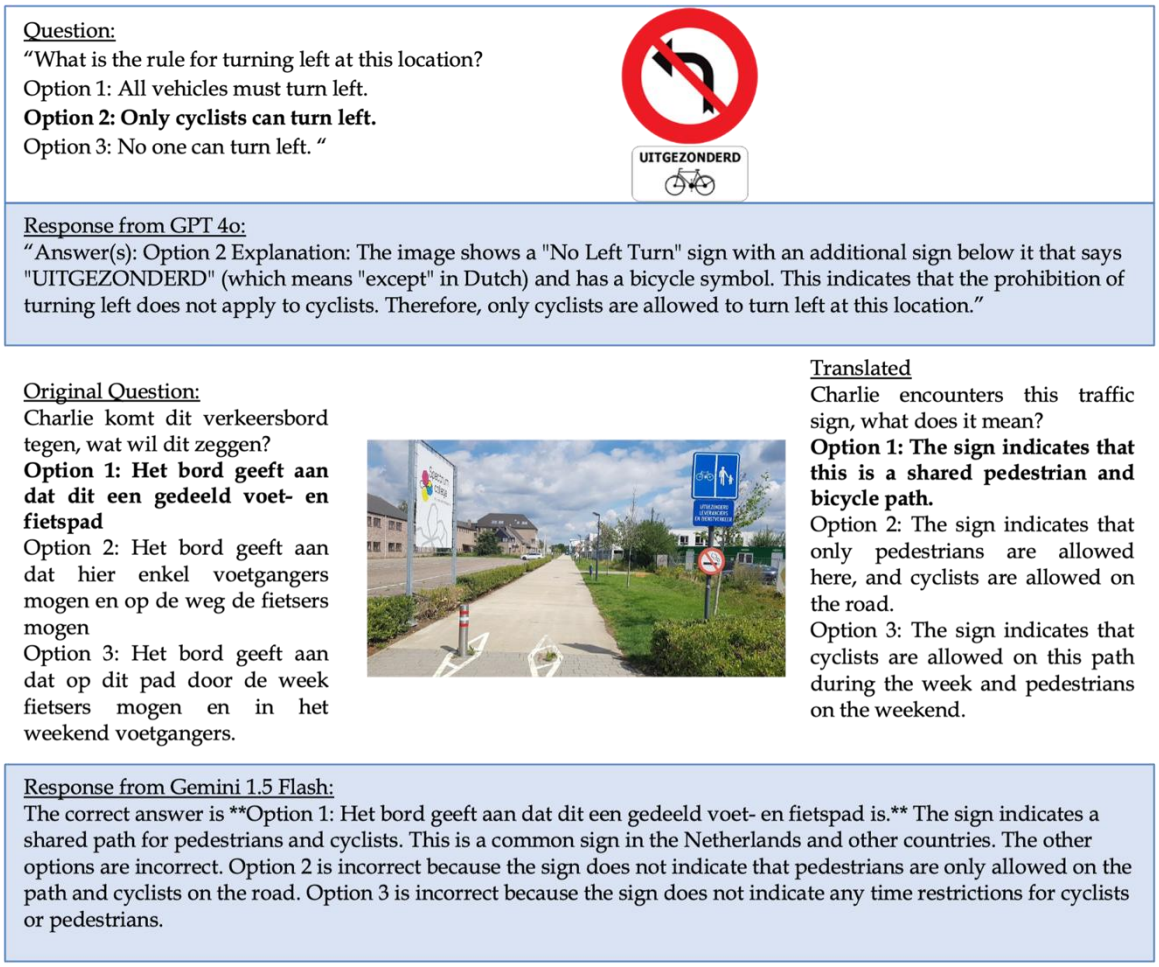


Figure 6. Examples of LLM responses, showing its capability in understanding text and interpreting images, including traffic boards and road layouts, to give a correct answer. Note: the correct option is marked in bold.

Although LLMs show the capacity to effectively process multilingual and multimodal inputs, the results indicate that LLMs perform better in English than in Dutch. When comparing zero-shot prompting on the BoolQ datasets for transportation understanding, content in Dutch had lower performance than English content by 18.6 percentage points in Gemini, by 12.1 percentage points in DeepSeek, and by 11 percentage points in GPT. GPT-4o, GPT-o1, and DeepSeek R1 outperformed the tested Gemini models. Gemini 1.5 Pro often refused to provide answers to questions in Dutch, citing difficulty in understanding the questions. Some example responses were: "I need this question translated from Dutch to English to understand it and provide an accurate answer. Could you please translate it?", or " I can't answer this question because it is written in Dutch and the model can only understand English.", or "I can't answer this question because it is in another language. I can translate English to Dutch, but I need the questions to be in English."

Other studies found similar trends: Lai et al. (2023) reported that ChatGPT performed better in English than in the other 36 languages examined. Ahuja et al. (2023) evaluated eight different tasks across 33 languages, concluding that GPT-3.5 turbo and GPT-4.5 turbo performed better in English

than in non-English languages. Another finding by Dey et al. (2024) showed that GPT-4 performed better than Gemini when using English prompts compared to prompts in South Asian languages. In the medical domain, GPT-4 also outperformed Gemini on Taiwan's Pulmonologist Exam, which featured non-English content and image-based questions (Chen et al., 2024).

5.3 Role of prompting

Instructional prompting was designed to guide the model toward producing consistent responses in the correct format. For most LLMs, the language of the instruction prompt, Dutch or English, yielded comparable results, although its effectiveness was influenced by the language of the content and the complexity of the task. Instructional prompting helped DeepSeek-R1 significantly in Dutch content in BoolQ_nl. However, it reduced the performance of Gemini models or had little impact on improving performance in GPT models in Boolean questions. The instructional prompt proved effective for multiple-choice answer questions (MulAns), which increased accuracy across all models. However, its relative impact was different on each model, demonstrated by F1, precision, and recall scores. Models with lower zero-shot baseline performance showed more benefit from instruction prompting than stronger-performing models. After applying instructional prompting, both types of models achieve relatively strong performance. At this stage, adding further complexity, such as few-shot examples in k-shot prompting, provides little additional benefit and can sometimes slightly reduce performance, as observed in both BoolQ datasets and MulAns datasets. This pattern suggests that instructional prompting effectively aligns models to the task, and once this alignment is achieved, extra complexity does not always yield improvements. Wang and Zhao (2023) observed that while extra context could be advantageous, it could also lead to inaccuracies, adding complexity to the prompt. Similarly, Liu et al. (2024) found that advanced prompting techniques, such as chain-of-thought prompting, could sometimes reduce performance in specific scenarios.

5.4 Human vs. LLM Performance

Both our study and the study by Riaz et al. (2019) used the same group of questions on the R2S platform, which allowed for a meaningful comparison between LLMs and children's performance. Specifically, the KidsKnowl dataset in our study corresponded to the traffic knowledge module, while the KidsRiskMgmt dataset aligned with the risk management module. Riaz et al. (2019) conducted their study with 44 elementary school pupils (ages 9–13) in Belgium. The modules were designed to include both familiar and unfamiliar situations based on where the images were captured and whether the location was familiar to the pupils' regions. According to their study, the pupils achieved an average of 93.41% in familiar situations and 87.05% in unfamiliar situations for traffic knowledge. For risk management, the mean scores were 91.82% in familiar situations and 88.41% in unfamiliar ones (Riaz et al., 2019).

Table 6. The performance of elementary school pupils in Flanders (Belgium)

Modules	Mean score - **familiar** situation	Mean score - **unfamiliar** situation
Knowledge	93.41	87.05
Risk management	91.82	88.41

* Adapted from "Evaluation of a Gamified E-Learning Platform to Improve Traffic Safety Among Elementary School Pupils in Belgium," by M. S. Riaz et al., 2019, *Personal and Ubiquitous Computing*, *23*(5), p. 935.

Compared to the children's performance, for KidsKnowl questions, GPT-4o slightly outperformed the children in unfamiliar situations (89.2% vs. 87.05%) but underperformed in familiar ones (89.2% vs. 93.41%) (Figure 4, Table 6). This comparison suggested that the model could recall traffic laws effectively in novel contexts. However, its performance in risk management tasks was significantly lower than that of the pupils across both familiar and unfamiliar scenarios. In the risk management tasks, it does not require only the understanding of traffic signs, but also requires more complex

skills to interpret the traffic situations. Children scored 91.82% and 88.41% in familiar and unfamiliar situations, respectively, whereas GPT-4o scored highest at 71.4% (Figure 4). It could indicate that while large language models like GPT-4o can handle direct questions involving recognition and recall of traffic rules, they may struggle with more context-dependent or indirect questions, the areas where children may apply reasoning based on experience and intuition.

One possible explanation is that, since the model is pre-trained on broad general knowledge, it may be less familiar with transportation-specific contexts or terminology. Providing more targeted "familiar context" through retrieval-augmented generation (RAG) or improving reasoning via techniques such as chain-of-thought (CoT) prompting could help address these gaps. Evidence from other domains supports the potential of these approaches. For example, Wei et al. (2022) showed that CoT prompting yielded substantial performance gains across several reasoning domains, including arithmetic, commonsense, and symbolic reasoning, particularly with the largest language models, and on complex problems. Significant gains have also been reported in specialized domains. In medical imaging, Parvaiz et al. (2024) demonstrated that structured reasoning improved caption accuracy nearly threefold in spine cancer image analysis. However, the effectiveness of CoT varies depending on the task. A large meta-analysis by Sprague et al. (2025), covering over 100 studies across 20 datasets and 14 models, found that CoT significantly improves mathematical and symbolic reasoning tasks but offers limited benefits for non-symbolic tasks such as commonsense reasoning or general language understanding.

Similarly, RAG has shown benefits in question-answering and knowledge-intensive tasks. Muludi et al. (2024) applied RAG to document-based question answering on the Stanford Question Answering Dataset and reported significant improvements over baseline systems. In addition, the study from Ram et al. (2023) found that adding in-context RAG without modifying LLM architecture allows smaller models to perform as well as much larger ones, making them more efficient. It also found that using simple retrievers to retrieve information significantly improves accuracy and reduces factual errors across different datasets. These findings suggest that future research could explore both CoT and RAG techniques to further enhance model performance.

6 Conclusion and future work

In this work, we evaluated the performance of the most recently released large language models, namely Gemini, GPT, and DeepSeek, across different prompting strategies, tasks, types of inputs, and languages to understand their multilingual capabilities and their expertise in transportation. We used global and local accuracy metrics to assess the models' ability to handle multiple-choice questions and single and multiple correct answers, including text and multimodal inputs.

Our investigation showed that pre-trained models performed well on transportation tasks, outperforming random guessing in both text-only and text-image scenarios. The performances varied among different models. For the more straightforward tasks, such as true-false questions in BoolQ datasets, the differences were not noticeable, but when the questions became more complex, like in MulAns, the performance gaps became more apparent among the tested models. In our evaluation on MulAns, GPT-o1 achieved the highest overall performance, outperforming both DeepSeek R1 and Gemini models. This suggests that a less powerful and cost-efficient model may suffice for simpler tasks, while more complex tasks may require a more advanced model like GPT-o1.

The language of the content affects LLM performance, with models generally performing better on English content than Dutch. Overall, GPT and DeepSeek models are less language-sensitive than Gemini. DeepSeek-R1 and GPT-4o consistently outperformed other models across various datasets and prompt settings, demonstrating robustness in handling both English and Dutch content effectively. They also showed superior performance in understanding and interpreting multimodal inputs. The Gemini models, especially Gemini 1.5 Pro, exhibited fluctuating performance and

sensitivity to language. Self-lingual and cross-lingual instruction prompting generally yield comparable results, indicating that cross-lingual IP can often substitute for self-lingual IP without major performance loss.

Our evaluation of the text-image dataset showed that large language models performed well above random guessing, suggesting they could interpret visual content and relate it to traffic regulations. GPT-4o performed comparably to young children in traffic knowledge questions, even slightly outperforming them in unfamiliar situations. However, it lagged behind children's performance in familiar contexts and performed significantly worse in risk management tasks, under the testing condition with 0-shot and instruction prompting. This may result from the pre-trained model's insufficient focus on transportation-specific contexts. Approaches like RAG or chain-of-thought CoT prompting could help bridge this gap by providing relevant information and reasoning guidance.

Depending on cost, context, and practicality, different LLMs can be selected. GPT-4o, GPT-o1 and DeepSeek-R-1 are less sensitive to language, making them particularly useful in multilingual contexts where questions may be mixed in Dutch and English. Under a zero-shot setting, they perform well for direct use and can serve as strong baselines for further fine-tuning. Smaller, lighter, and cheaper models (Gemini 1.5 flash, GPT-3.5-Turbo) can also achieve good results, particularly in English, with appropriate instructional prompting. Intermediate steps may further improve outcomes, including improved prompt design, the use of CoT prompting to enhance reasoning, the incorporation of RAG to integrate local transportation knowledge, the translation of user requests, and the translation of databases used for RAG…

Overall, this study provides insights into the baseline performance of various LLMs, serving as guidance for selecting models for specific transportation applications and for subsequent improvement efforts.

6.1 Limitations and Future Work

The study is limited by the set of models evaluated, as new models continue to emerge in the rapidly evolving LLM ecosystem. Due to practical constraints, multimodal evaluation did not include DeepSeek, whereas models from OpenAI's GPT family and Google's Gemini family were considered. Future work could expand the evaluation to include a broader range of LLMs for a more comprehensive comparison. Our study also assessed a limited number of datasets and scenarios in transportation; we plan to scale up our evaluation in the future to expand the study. The findings show the potential of using LLMs in transportation applications, with considerable room for performance improvement. The inconsistent performance of some models suggests the need for further fine-tuning and domain-specific adjustments, including developing specialized models tailored to specific transportation tasks and contexts or RAG application to provide extra knowledge for the pre-trained LLMs. The performance of LLMs could be sensitive to prompting; however, we did not conduct extensive prompt tuning or experiment with alternative prompting designs, such as CoT, due to practical time limitations. Future work will explore this limitation further. In addition, future work could further investigate the application of LLMs in transportation, particularly by integrating LLMs enhanced with RAG and CoT techniques into a human-in-the-loop decision-support system to support transportation experts in the Dutch context.

Funding

This work was supported by the Flemish government (under the project "STRIKE: SubjecTieve veRkeersveiligheid Innovatief aanpaKken en vErbeteren").

Data Access Statement

The data used in this study are confidential and cannot be publicly shared due to their ongoing use in educational programs and commercial products. They may be shared upon request, subject to applicable restrictions.

Contributor Statement

Thi M.D. Tran: Conceptualization, data curation, formal analysis, methodology, validation, writing-original draft preparation, writing-reviewing and editing. Davy Janssens: Conceptualization, methodology, funding acquisition. Geert Wets: Conceptualization, methodology, funding acquisition, resources. Tom Brijs: Conceptualization, methodology, funding acquisition. Burcu Can: Resources, validation, writing-reviewing and editing. Jan Vuurstaek: Data curation. Lien Aerts: Resources, data curation. Wim Ectors: Conceptualization, methodology, formal analysis, validation, writing-reviewing and editing.

Use of AI

During the preparation of this work, the authors used Grammarly to improve and correct the grammar and spelling of this work during its preparation. After using this tool, the authors reviewed, edited, made the content their own, and validated the outcome as needed, and take full responsibility for the content of the publication.

Acknowledgements

We thank our colleagues who supplied the questions we used in our study: Prof. dr. Ariane Cuenen, Roeland Paul, Dr. Hélène Dirix, Prof. dr. Kris Brijs, and Prof. dr. Evelien Polders. We also appreciate Dirk Roox for retrieving data from the online platform. Additionally, we thank An Langenaken and Rune Rombouts for their assistance with data annotation.

Conflict Of Interest (COI)

There is no conflict of interest.

References

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., & Sitaram, S. (2023). MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 4232–4267). https://doi.org/10.18653/v1/2023.emnlp-main.258

Alghisi, S., Rizzoli, M., Roccabruna, G., Mousavi, S. M., & Riccardi, G. (2024). Should we fine-tune or RAG? Evaluating different techniques to adapt LLMs for dialogue. In S. Mahamood, N. L. Minh, & D. Ippolito (Eds.), *Proceedings of the 17th International Natural Language Generation Conference* (pp. 180–197). https://doi.org/10.18653/v1/2024.inlg-main.15

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science, 8,* 17. https://doi.org/10.1007/s42001-024-00345-9

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Nguyen, V. Q., Do, T. T., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, *1*, 675–718. https://doi.org/10.18653/v1/2023.ijcnlp-main.45

Buscaldi, D., & Rosso, P. (2023). How good is NLLB for low-resource languages? A study on the Genoese language. *CEUR Workshop Proceedings*, *3596*. https://ceur-ws.org/Vol-3596/short5.pdf

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P., & Sun, L. (2025). A survey of AI-generated content (AIGC). *ACM Computing Surveys*, *57*(5), 1–38. https://doi.org/10.1145/3704262

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Chen, C., Hsieh, K., Huang, K., & Lai, H. (2024). Comparing vision-capable models, GPT-4 and Gemini, with GPT-3.5 on Taiwan's pulmonologist exam. *Cureus*, *16*(8), e67641. https://doi.org/10.7759/cureus.67641

Common Crawl Foundation. (n.d.). *Statistics of Common Crawl monthly archives by commoncrawl*. Retrieved July 29, 2024, from https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health *Frontiers in Public Health*, *11*, Article 1166120. https://doi.org/10.3389/fpubh.2023.1166120

Deep Infra. (2025). *DeepSeek V-3*. Retrieved April 23, 2025, from https://deepinfra.com/deepseek-ai/DeepSeek-V3

DeepSeek API Docs. (2025). *DeepSeek-R1 Release*. Retrieved April 22, 2025, from https://api-docs.deepseek.com/news/news250120

DeepSeek. (2025). *DeepSeek AI: chat, models & independent guides*. Retrieved April 20, 2025, from https://deepseek.ai/

Dey, K., Tarannum, P., Hasan, M. A., Razzak, I., & Naseem, U. (2024). *Better to ask in English: Evaluation of large language models on English, low-resource and cross-lingual settings.* arXiv. https://doi.org/10.48550/arXiv.2410.13153

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71, Article 102642*. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., Hemphill, L. (2024). A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, *15*(5), 1–25. https://doi.org/10.1145/3664930

Gemini Team. (2023). *Gemini: A family of highly capable multimodal models*. arXiv. https://doi.org/10.48550/arXiv.2312.11805

Guo, X., Zhang, Q., Jiang, J., Peng, M., Zhu, M., Yang, H.F. (2024). Towards explainable traffic flow prediction with large language models. *Communications in Transportation Research, 4*(100150). https://doi.org/10.1016/j.commtr.2024.100150

Hammarström, H. (2016). Linguistic diversity and language evolution. *Journal of Language Evolution, 1*(1), 19–29. https://doi.org/10.1093/jole/lzw002

Knill, K., & Young, S. (1997). Hidden Markov models in speech and language processing. In S. Young & G. Bloothooft (Eds.), *Corpus-based methods in language and speech processing* (pp. 27–68). Springer. https://doi.org/10.1007/978-94-017-1183-8_2

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 22199–22213). Neural Information Processing Systems Foundation. https://doi.org/10.52202/068431-1613

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., Nguyen, T. (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *In Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 13171-13189). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Langhe, L.D., Maladry, A., Vanroy, B., Bruyne, L.D., Singh, P., Lefever, E., Clercq, O.D. (2024). Benchmarking Zero-Shot Text Classification for Dutch. *Computational Linguistics in the Netherlands, 13, 63–90.* https://www.clinjournal.org/clinj/article/view/172

Li, Y., Li, L., Wu, Z., Bing, Z., Xuanyuan, Z., Knoll, A.C., Chen, L. (2024). UnstrPrompt: Large Language Model Prompt for Driving in Unstructured Scenarios. *IEEE Journal of Radio Frequency Identification, 8, 367-375*. https://doi.org/10.1109/JRFID.2024.3367975

Liang, Y., Liu, Y., Wang, X., Zhao, Z. (2024). Exploring large language models for human mobility prediction under public events. *Computers, Environment and Urban Systems 112*(102153). https://doi.org/10.1016/j.compenvurbsys.2024.102153

Lin, X.V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P.S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., Stoyanov, V., Li, X. (2022). Few-shot Learning with Multilingual Generative Language Models. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9019-9052). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.616

Liu, R., Geng, J., Wu, A.J., Sucholutsky, I., Lombrozo, T., Griffiths, T.L. (2024). *Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse*. arXiv. https://doi.org/10.48550/ARXIV.2410.21333

Mo, B., Xu, H., Zhuang, D., Ma, R., Guo, X., Zhao, J. (2023). *Large Language Models for Travel Behavior Prediction*. arXiv. https://doi.org/10.48550/arXiv.2312.00819

Muller, B., Anastasopoulos, A., Sagot, B., Seddah, D. (2021). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 448-462). Association for Computational Linguistics https://doi.org/10.18653/v1/2021.naacl-main.38

Muludi, K., Fitria, K.M., Triloka, J., Sutedi. (2024). Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. *International Journal of Advanced Computer Science and Applications, 15(3)* . https://doi.org/10.14569/IJACSA.2024.0150379

Nazi, Z.A., Hossain, Md.R., Mamun, F.A. (2025). Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal, 10*(100124). https://doi.org/10.1016/j.nlp.2024.100124

OpenAI. (2024a). *Hello GPT-4o*. Retrieved July 30, 2024. https://openai.com/index/hello-gpt-4o/

OpenAI. (2024b). *GPT-4 Technical.* arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2022a). *Introducing ChatGPT*. Retrieved June 3, 2024. https://openai.com/blog/chatgpt

OpenAI. (2022b). *DALL·E 2*. Retrieved July 31, 2024. https://openai.com/index/dall-e-2/

Parvaiz, A., Nasir, E.S., Fraz, M.M. (2024). From Pixels to Prognosis: A Survey on AI-Driven Cancer Patient Survival Prediction Using Digital Histology Images. *Journal of Imaging Informatics in Medicine, 37, 1728–1751.* https://doi.org/10.1007/s10278-024-01049-2

Pichai, S., Hassabis, D. (2024). *Our next-generation model: Gemini 1.5*. Google. Revised July 3, 2024. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y. (2023). In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics, 11, 1316–1331.* https://doi.org/10.1162/tacl_a_00605

Ray, P.P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3, 121–154.* https://doi.org/10.1016/j.iotcps.2023.04.003

Reynolds, D. (2009). Gaussian Mixture Models. In Li, S.Z., Jain, A. (Eds.*) Encyclopedia of Biometrics* (pp. 659-663). Springer. https://doi.org/10.1007/978-0-387-73003-5_196

Riaz, M.S., Cuenen, A., Janssens, D., Brijs, K., Wets, G. (2019). Evaluation of a gamified e-learning platform to improve traffic safety among elementary school pupils in Belgium. *Personal and Ubiquitous Computing, 23, 931–941*. https://doi.org/10.1007/s00779-019-01221-4

Sprague, Z.R., Yin, F., Rodriguez, J.D., Jiang, D., Wadhwa, M., Singhal, P., Zhao, X., Ye, X., Mahowald, K., Durrett, G. (2025). To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. *In Proceedings of the International Conference on Learning Representations (ICLR 2025)* (pp. 94118–94162). https://doi.org/10.48550/arXiv.2409.12183

Taal Unie. (n.d.). *Feiten & cijfers – wat iedereen zou moeten weten over het Nederlands*. Retrieved July 29, 2024. https://taalunie.org/informatie/24/feiten-cijfers

Vanroy, B. (2023). *Language Resources for Dutch Large Language.* arXiv. https://doi.org/10.48550/arXiv.2312.12852

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Łukasz, Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.) *Advances in Neural Information Processing Systems* (pp. 6000-6010). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1706.03762

Wang, Y., Zhao, Y. (2023). *Gemini in Reasoning: Unveiling Commonsense in Multimodal Large Language Models*. arXiv. https://doi.org/10.48550/ARXIV.2312.17661

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V., Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (pp. 24824-24837). Neural Information Processing Systems Foundation. https://doi.org/10.52202/068431-1800

Yan, H., Li, Y. (2025). Generative AI for Intelligent Transportation Systems: Road Transportation Perspective. *ACM Computing Surveys, 57*(12), *1-45.* https://doi.org/10.1145/3719290

Zhang, S., Fu, D., Zhang, Z., Yu, B., Cai, P. (2023). *TrafficGPT: Viewing, Processing and Interacting with Traffic Foundation Models.* arXiv. https://doi.org/10.48550/ARXIV.2309.06719

Zhang, Z., Sun, Y., Wang, Z., Nie, Y., Ma, X., Li, R., Sun, P., Ban, X. (2025). *Large Language Models for Mobility Analysis in Transportation Systems: A Survey on Forecasting Tasks*. arXiv. https://doi.org/10.48550/ARXIV.2405.02357