Performance of Large Language Models in Domain-Specific and Underrepresented Languages: A Case Study on the Transportation Domain and Dutch Language

Authors: 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 (IMOB), UHasselt, Martelarenlaan 42, 3500 Hasselt, Belgium

[2] Computing Science and Mathematics, University of Stirling, Stirling, UK, FK9 4LA

UHASSELT KNOWLEDGE IN ACTION

INTRODUCTION

Motivations

- LLMs excel in general tasks, primarily in English, but their performance in domain-specific reasoning and underrepresented languages (like Dutch) remains underexplored.
- Cross-lingual capabilities in specialized domains have not been widely studied.

Objectives

- Enhance understanding of cross-lingual capabilities in specialized domains.
- Explore transfer learning potential for underrepresented languages like Dutch.
- Aid in selecting effective LLM foundation models for domain-specific applications.
- Provide performance benchmarks for LLMs in Dutch for transportation tasks.

EXPERIMENTAL SETUP

- Question–answer format, extracted from our teaching and training materials at the School of Transportation Sciences and the Transportation Research Institute, UHasselt, Belgium
- 991 questions distributed across six datasets
- Include text only, text & image

MulAns (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

Instructional Prompting (IP_en)

The following question is under the transportation context and regulations in Belgium:
Instruction:
- Indicate the correct option number in your answer, beginning with "Answer: " such as "Answer: Option 1"
- Answer in English.
The question is:

BoolQ_en

By using the engine brake, the speed is reduced more smoothly, and you save the brake pads.
Option 1: True
Option 2: False

		Datasets					
		BoolQ_en	BoolQ_nl	MulAns	BasicKnowl	KidsKnowl	KidsRiskMgmt
Data	Nr of questions	270	227	188	50	130	126
	Language	English	Dutch	Dutch	English	Dutch	Dutch
	Type	text	text	text	text & image	text & image	text & image
LLM	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
Prompt settings	0-shot SP	x	x	x	x	x	x
	IP_en	x	x	x	x	x	x
	IP_nl	x	x				
	k-shot & IP_en	x	x	x			

K-shot & IP_en

The following question is under the transportation context and regulations in Belgium:
Instruction:
- Indicate the correct option number in your answer, beginning with "Answer: " such as "Answer: Option 1"
- Answer in English.
Examples:
"The question is: True or false, speed boards display the recommended travel speed for a line, it is not illegal to drive quicker than the posted speed? ['Option 1:True', 'Option 2:False']
Answer: Option 1"
"The question is: Sleep is the only effective remedy for sleepiness. ['Option 1:True', 'Option 2:False']
Answer: Option 1"
The question is:

KidsKnowl (translated)

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 (translated)

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.

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.

EVALUATION METHOD

$$global\ accurancy\ (GA)\ (\%) = \frac{total\ correct\ responses}{total\ questions} x100$$

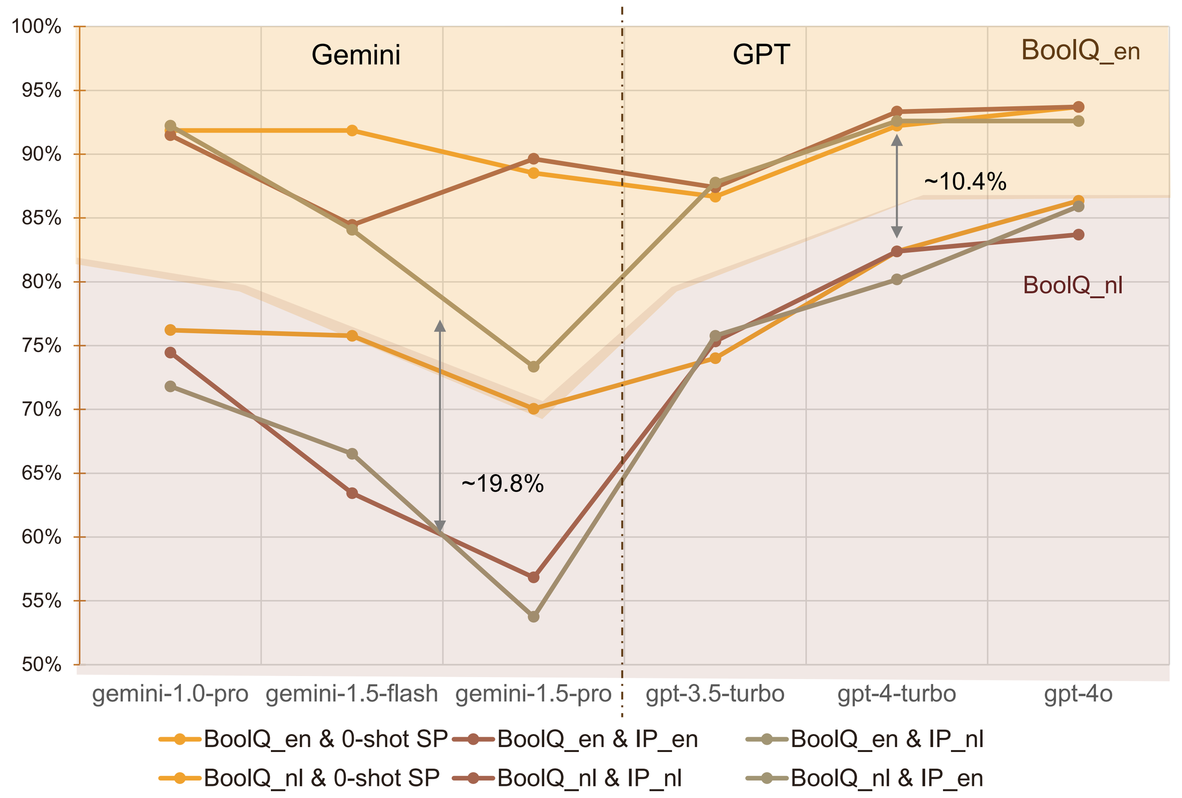
$$local\ accurancy\ (LC)\ (\%) = \frac{\sum nr\ of\ correct\ indices\ per\ response}{\sum nr\ of\ indices\ per\ question} x100$$

<u>Example Question:</u>
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.
<u>Response from an LLM:</u> "The correct answer is Option 3"
- *Ground truth: [1,0,1]; LLM's answer: [0,0,1].*
- *The global accuracy is 0%*
- *The local accuracy 60%*

FINDINGS

Impact of Language on LLM's Performance in the Transportation Domain

- LLM generally performs better with English than Dutch content.
- Language performance gap varies
- GPT models handle different content languages better than Gemini models
- Average accuracy variance: GPT models (~10.4%) vs. Gemini models (~19.8%).
- Language of the instruction plays a small impact



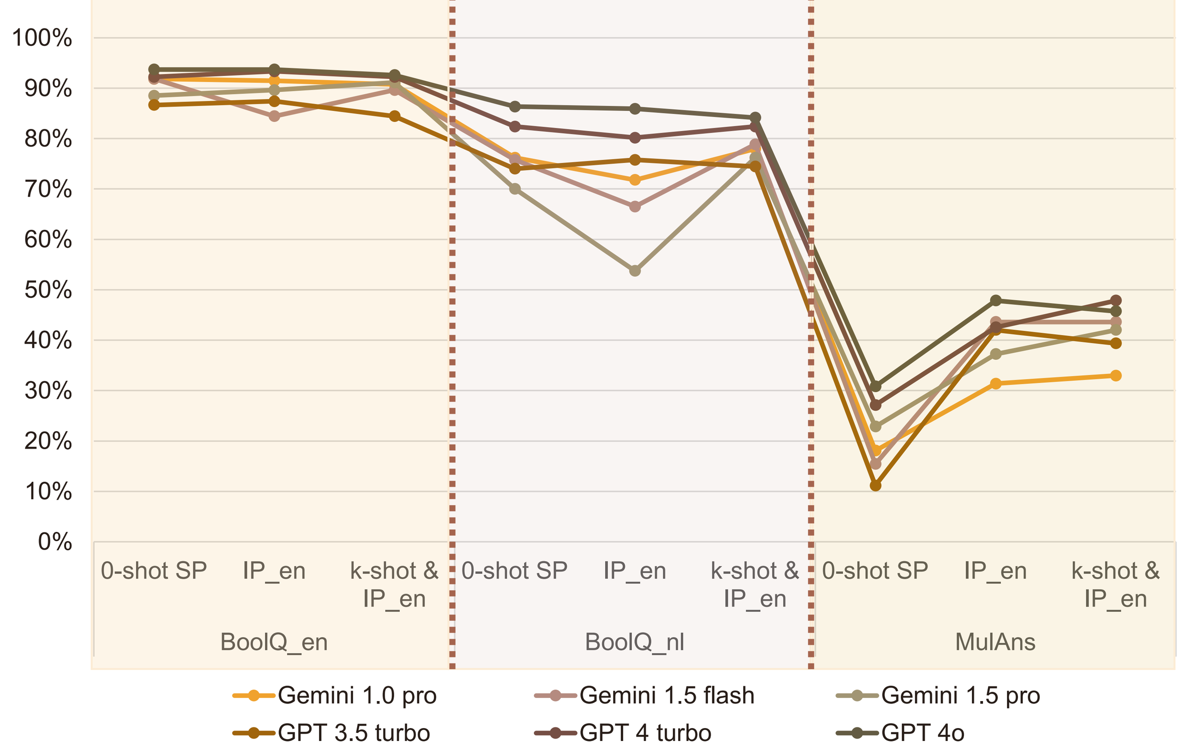
Performance comparison of six LLMs across two different languages in content and instructional prompt (GA)

Impact of Few-Shot Prompting

Mixed results:
- **Gemini models:** Performance improves when examples follow instructions
- **GPT models:** accuracy for GPT-4 turbo

Impact of Instructional Prompting

- Enhances accuracy for multiple-answer questions across all models.
- Useful in handling complex tasks



Performance comparison of six LLMs across different text datasets (GA)

Performance Differences Between Question Complexity

- **Multiple-answer questions:**
  - Require complex understanding, logical reasoning, and problem-solving.
  - Improved with instructional prompting.
- **Boolean questions:**
  - Simpler due to its binary nature
  - Instructional prompting has minimal impact

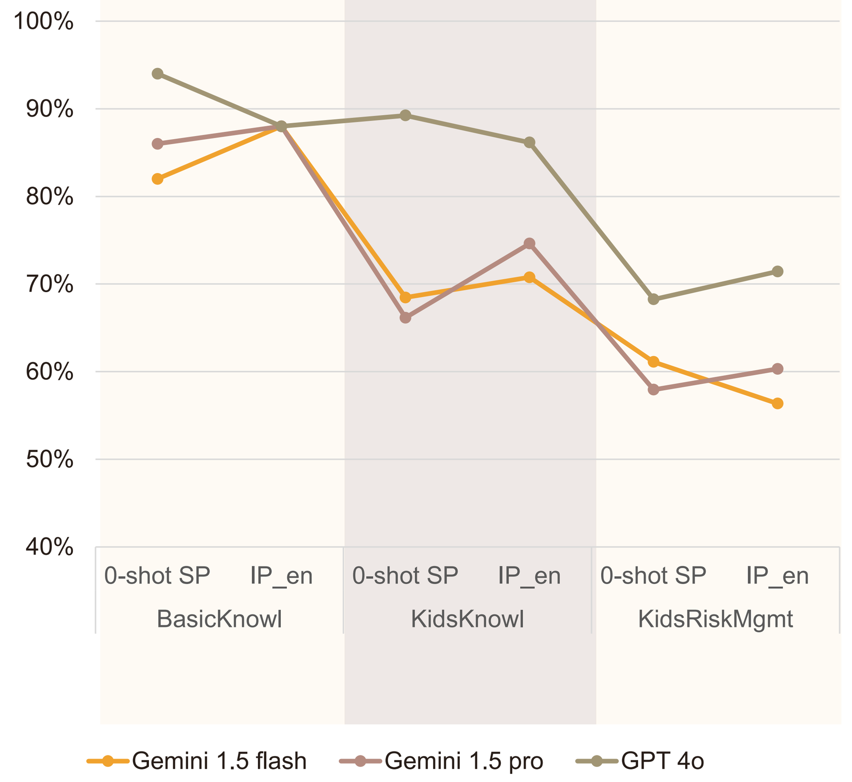
Models	0-shot SP		IP_en		k-shot & IP_en	
	GA	LA	GA	LA	GA	LA
Gemini-1.0-pro	18.09%	64.55%	31.38%	73.21%	32.98%	72.75%
Gemini-1.5-flash	15.43%	65.94%	43.62%	77.48%	43.62%	78.41%
Gemini-1.5-pro	22.87%	64.67%	37.23%	72.86%	42.02%	77.37%
GPT-3.5-turbo	11.17%	63.86%	42.02%	77.83%	39.36%	76.33%
GPT-4-turbo	27.13%	72.17%	42.55%	80.14%	47.87%	80.37%
GPT-4o	30.85%	74.13%	47.87%	82.22%	45.74%	83.03%

Global accuracy (GA) and local accuracy (LA) of MulAns dataset

- Local accuracy aligns with global accuracy
- Performance above random guessing (>50% LA)
- The gap between GP & LA's → potential for improvement with IP

Performance on Text and Image-Based Transportation Tasks

- Better than random.
- **GPT-4o** consistently outperformed **Gemini models**
- Better performance with common knowledge (BasicKnowl).
- Decreased performance with specialized knowledge (KidsKnowl).
- Lowest performance on complex, domain-specific tasks (KidsRiskMgmt)



Performance comparison of three LLMs across different multimodal input datasets

CONCLUSION

Performance on Transportation Tasks:

- outperformed random guessing in both text-only and text-image scenarios.

Language Sensitivity:

- Performed better with English content than Dutch.
- Less sensitivity to language differences of GPT's than Gemini's.

Model Comparison:

- GPT-4o consistently.
- Gemini models, (Gemini 1.5 Pro): fluctuating performance and higher sensitivity to language.

Implications:

- Provides a deeper understanding of LLM performance in transportation tasks, especially in Dutch.
- Offers valuable insights for selecting a suitable LLM for tasks involving specialized domains and underrepresented languages.

LIMITATIONS

- Evaluated six Gemini and GPT models; excluded open-source multimodal LLMs
- Due to closed source, limited insights on the LLM architecture, flexibility
- Limited number of datasets and transportation scenarios.

FUTURE WORK

- Expand the scope of evaluation to include a broader range of datasets and scenarios in transportation.
- Incorporate open-source multimodal LLMs
- Improve performances by fine-tuning in the Dutch language, fine-tuning LLMs tailored for specialized transportation tasks and contexts, optimizing prompting

ACKNOWLEDGMENTS

- The Flemish government for funding this project
- Support from colleagues for question acquisition and data annotation.