Language and Planning in Robotic Navigation: A Multilingual Evaluation of State-of-the-Art Models

Anonymous submission

Abstract

Large Language Models (LLMs) such as GPT-4, trained on huge amount of datasets spanning multiple domains, exhibit significant reasoning, understanding, and planning capabilities across various tasks. This study presents the first-ever work in Arabic language integration within the Vision-and-Language Navigation (VLN) domain in robotics, an area that has been notably underexplored in existing research. We perform a comprehensive evaluation of state-of-the-art multilingual Small Language Models (SLMs), including GPT-4o mini, Llama 3 8B, and Phi-3 medium 14B, alongside the Arabic-centric LLM, Jais. Our approach utilizes the NavGPT framework, a pure LLM-based instruction-following navigation agent, to assess the impact of language on navigation reasoning through zero-shot sequential action prediction using the R2R dataset.

Through comprehensive experiments, we demonstrate that our framework is capable of high-level planning for navigation tasks when provided with instructions in both English and Arabic. However, certain models struggled in reasoning and planning in the Arabic language due to limitations in their reasoning capabilities, poor performance, and parsing issues. These findings highlight the importance of enhancing reasoning capabilities in language models for effective navigation, emphasizing this as a key area for further development, while also unlocking the potential of Arabic-language models for impactful real-world applications.

Keywords *Vision-and-Language Navigation (VLN), Small Language Models (SLMs), Large Language Models (LLMs), Jais, NavGPT, Reasoning, Robotics*

Introduction

With the rise of AI-driven robotics in smart cities, multilingual interaction systems have become increasingly crucial, particularly in the Middle East and North Africa (MENA) region, where investment in autonomous systems is growing rapidly. However, the scarcity of models trained on Arabic data presents a significant barrier to their deployment. Despite Arabic's importance, spoken by over 400 million people (Koto et al. 2024), its under-representation in Vision-and-Language Navigation (VLN) research limits the effectiveness of autonomous systems in Arabic-speaking regions. This research addresses these gaps by examining how Arabic and English inputs influence the performance of Language Models (LMs) in robotic VLN tasks. The goal is to contribute to developing more inclusive autonomous systems that address the linguistic and cultural diversity of the Arabic-speaking world by seamlessly understanding and executing instructions in Arabic.

Small Language Models (SLMs) are more efficient and cost-effective than Large Language Models (LLMs) due to their smaller size and reduced computational requirements, making them suitable for deployment on resource-constrained devices (Lu et al. 2024). Their smaller footprint enables faster response times and easier integration into existing systems. Despite these advantages, SLMs have received comparatively less attention in research, presenting an opportunity to explore their potential for optimization, efficiency, and domain-specific applications.

In this work, we focus on SLMs rather than Vision-Language Models (VLMs) due to SLMs' superior ability to process and reason about complex linguistic structures across multiple languages (Lu et al. 2024). As noted by (Rahmanzadehgervi et al. 2024), VLMs often face performance limitations in real-world navigation tasks, particularly because their approach to extracting visual features tends to neglect instruction prompts, thereby reducing adaptability. Moreover, excluding vision from the reasoning process mitigates potential simulation-to-reality gaps during robot deployment.

Our evaluation involves inferencing and comparing state-of-the-art multilingual SLMs and Core42's Arabic-centered LLM, Jais 30B (Sengupta et al. 2023), on the NavGPT framework. The SLMs are OpenAI's GPT-4o mini (Vidhyashree 2024), Meta's Llama 3 8B (Dubey 2024), and Microsoft's Phi-3 medium 14B (Abdin et al. 2024). Using the Room-to-Room (R2R)-VLN dataset (Anderson et al. 2018), which provides English-language navigation instructions, we augment the data with Arabic translations generated using the Groq API for comparative analysis. The models are evaluated within the NavGPT framework, which operates in a zero-shot manner to predict sequential actions based on textual descriptions of visual observations, navigation history, and navigable viewpoints as is shown in Figure 1 (Zhou, Hong, and Wu 2023).

Contributions

The contributions of this work are as follows:

- Developed the first-ever framework incorporating Arabic

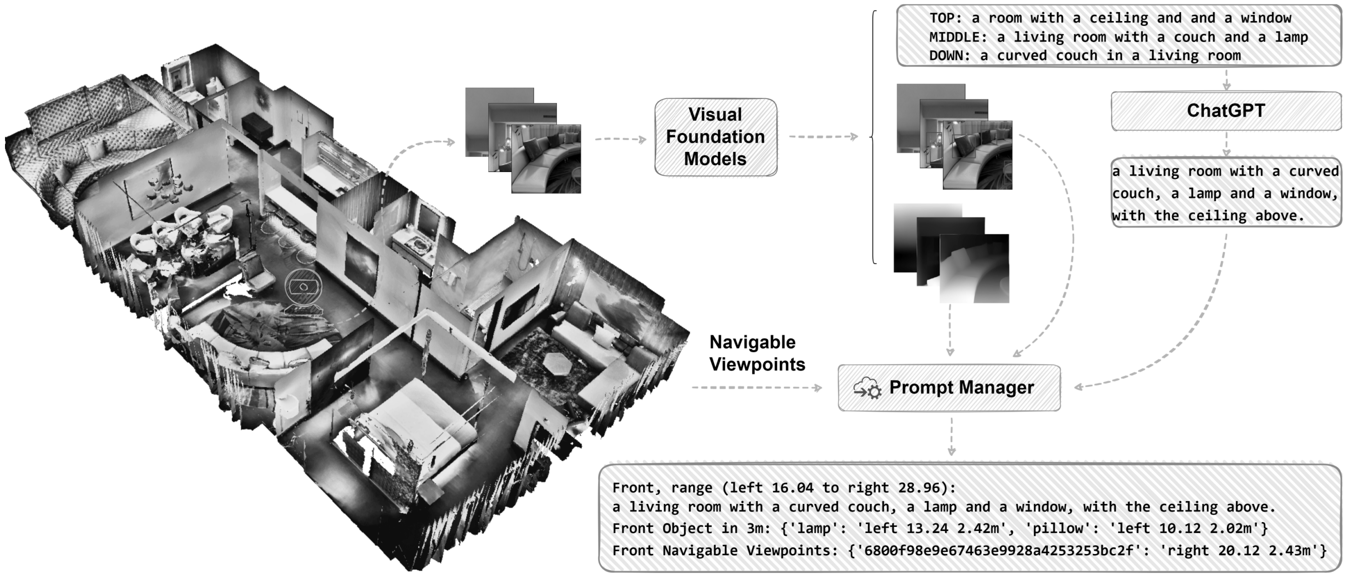

Figure 1: NavGPT methodology diagram (Zhou, Hong, and Wu 2023)

instructions in the VLN problem to evaluate the performance of multilingual SLMs (GPT-4o mini, Llama 3 8B, Phi-3 medium 14B) and the Jais 30B LLM on VLN tasks using both Arabic and English instructions.

- Augmented the R2R-VLN dataset with Arabic translations via the Groq API, creating a bilingual dataset for VLN research to address linguistic diversity.
- Conducted an in-depth analysis of the impact of instruction language on the reasoning capabilities of LMs.
- Identified key insights into language-specific limitations and strengths in multilingual SLMs and the Jais LLM, informing the development of autonomous systems suited for Arabic-speaking regions.

Related Works

VLN Tasks

VLN tasks use natural language instructions to guide agents through diverse environments, testing the reasoning and comprehension capabilities of LMs (Gu et al. 2022). These tasks typically evaluate agents in indoor settings, such as navigating through rooms in a building (Anderson et al. 2018), or outdoor scenarios, like traversing open landscapes or urban areas (Vargas-Munoz et al. 2021). Models must interpret navigation commands, align them with visual context, and generate appropriate actions in these varied environments.

Despite their promise, VLN tasks pose significant challenges. Ambiguities in language instructions, such as vague descriptions or inconsistent phrasing, can hinder accurate alignment with visual cues (Anderson et al. 2018). Sparse or incomplete visual data, such as in dim indoor settings or cluttered outdoor scenes, complicate navigation (Rahmanzadehgervi et al. 2024). These challenges are exacerbated when operating in multilingual contexts, where instructions may carry linguistic nuances that models struggle to capture or disambiguate. Addressing these limitations is crucial for improving real-world deployment of VLN systems (Ku et al. 2020; Tan, Yu, and Bansal 2019).

Language Models in Navigation

LMs play a pivotal role in VLN tasks by interpreting instructions and guiding agents through diverse environments. SLMs, LLMs, and VLMs each have distinct capabilities. SLMs are lightweight and efficient, enabling real-time processing in resource-constrained settings (Lu et al. 2024). Their streamlined architecture ensures practical deployment, though they may lack the extensive training and generalization capabilities of LLMs. By contrast, LLMs excel in multilingual reasoning and zero-shot tasks due to their vast training on diverse datasets (Zhou, Hong, and Wu 2023; Zhou et al. 2024). However, their computational demands and reliance on text-only reasoning limit their suitability for scenarios requiring multimodal integration (Lu et al. 2019).

VLMs specialize in combining visual and textual inputs, excelling in spatial reasoning tasks where strong visual grounding is critical. Despite this, VLMs often underperform in linguistically complex or multilingual scenarios, as their architectures prioritize visual processing over comprehensive language understanding (Rahmanzadehgervi et al. 2024; Koto et al. 2024). This study focuses on SLMs' ability to handle linguistic challenges like Arabic's morphological richness and syntactic complexity. This can enhance model robustness through exposure to diverse linguistic structures while ensuring efficient deployment in real-world applications, particularly in resource-constrained environments common in the MENA region.

Multilingual Challenges and Arabic-Specific Context

While most VLN studies focus on English-language instructions, they overlook the complexities and opportunities of non-English languages (Zhang et al. 2024). This work explores multilingual capabilities, particularly Arabic. Arabic, a widely spoken but underrepresented language in VLN, introduces unique challenges due to its morphological richness, syntactic complexity, and right-to-left script (Khalati, Ali, and Al-Romany 2024). These linguistic features make Natural Language Processing (NLP) tasks, including VLN, more demanding. Furthermore, the scarcity of high-quality Arabic datasets exacerbates these challenges, limiting the performance and adaptability of existing models (Ku et al. 2020).

Arabic-centric models like Jais (Sengupta et al. 2023) provide a foundation for addressing these gaps by leveraging training data tailored to Arabic. However, many VLN datasets, such as R2R (Anderson et al. 2018), remain focused on English, requiring augmentation or translation to support multilingual research. This study tackles these issues by analyzing how Arabic-language instructions impact reasoning in state-of-the-art models and identifying areas where existing architectures fall short in supporting Arabic tasks.

Datasets and Simulation Environments

VLN research relies heavily on paired visual and linguistic datasets. The R2R dataset (Anderson et al. 2018), with its English-language instructions for navigating photo-realistic environments, serves as a standard benchmark for evaluating VLN models' spatial reasoning and language comprehension. Simulation environments, such as Matterport3D (Chang et al. 2017), are crucial for testing models in realistic indoor settings, providing diverse and visually rich scenarios for evaluating navigation performance. Building on R2R, the RxR dataset (Ku et al. 2020) introduces multilingual instructions, including languages like Hindi, but excludes Arabic, which facilitates research on cross-linguistic adaptability.

Currently, Arabic navigation datasets are scarce. OpenStreetMap (Vargas-Munoz et al. 2021), a multilingual outdoor navigation dataset, includes limited non-English instructions, though Arabic coverage remains minimal. Given these limitations, our study leverages a translated R2R dataset to compare English and Arabic reasoning, highlighting language effects on navigation accuracy and providing groundwork for broader multilingual VLN research.

Problem Statement

The MENA region's growing reliance on autonomous systems emphasizes the need for multilingual VLN models capable of understanding Arabic, a language spoken by over 400 million people (Koto et al. 2024). Arabic's morphological richness and syntactic complexity present unique challenges for NLP (Khalati, Ali, and Al-Romany 2024), making it a valuable test case for evaluating language model reasoning. Despite this importance, existing VLN datasets, such as R2R, are predominantly in English (Anderson et al. 2018), leaving a critical gap in resources for Arabic-language tasks.

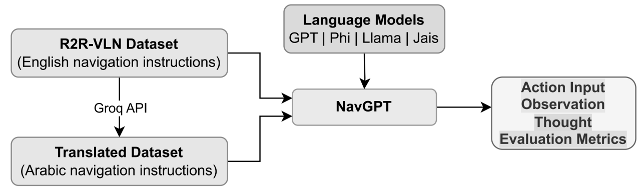

Figure 2: Evaluation pipeline

LLMs like GPT (OpenAI 2024) and Jais (Sengupta et al. 2023) have demonstrated strong multilingual reasoning capabilities, but their performance in Arabic navigation contexts remains under-explored. SLMs are cost-effective and practical for deployment on resource-constrained devices, such as for real-time navigation; however, their potential remains underrepresented in VLN research. Few studies have compared SLMs to LLMs in the context of multilingual navigation, highlighting a gap in the literature (Vidhyashree 2024). Addressing these gaps, this study explores how Arabic and English instructions affect reasoning capabilities in state-of-the-art SLMs and LLMs, aiming to improve inclusivity and adaptability in robotic navigation systems.

Methodology

In this section, we outline the problem formulation, method for translating the English R2R dataset to Arabic, and how NavGPT, a state-of-the-art LM-based navigation system, is adapted to process Arabic instructions. Our approach encompasses dataset translation, model setup, inference execution, and performance evaluation laid out in the pipeline in Figure 2.

VLN Problem Formulation

NavGPT addresses the VLN problem by framing it as follows (Zhou, Hong, and Wu 2023). Given a natural language instruction W, represented as a sequence of words $w_1, w_2, w_3, \ldots, w_{nu}$, the agent retrieves an observation O at each step s_t by interpreting its current location through a simulator. This observation consists of N alternative viewpoints, representing the surrounding environment of the agent in varying angles.

Each viewpoint observation is denoted as $o_i^p \in R$, with its corresponding angle direction represented as $a_i^p \in R$. Consequently, the observation at step t can be expressed as:

$$O_t = \left((o_1, a_1), (o_2, a_2), \ldots, (o_N, a_N) \right)$$

During navigation, the agent's action space is restricted to the navigation graph G. At each step, the agent selects the next action from the $M = |C_{t+1}|$ set of navigable viewpoints, C_{t+1}. This selection is guided by aligning the current observation O_{C_t} with the provided instruction W. The agent predicts the next action by identifying the relative angle a_{C_i} from O_{C_t}, executes this action through interaction with the simulator, and transitions from the current state $s_t = (v_t, \theta_t, \phi_t)$ to the next state $s_{t+1} = (v_{t+1}, \theta_{t+1}, \phi_{t+1})$, where v, θ, and ϕ represent the agent's current viewpoint, heading, and elevation angle, respectively.

To support navigation, the agent maintains a history of its previous states h_t and updates the conditional transition probability between states as follows:

$$S_t = T\left(s_{t+1} | a_{C_i}, s_t, h_t\right)$$

where T represents the conditional transition probability distribution.

In summary, the agent learns a policy π parametrized by Θ that relies on the instruction W and the current observation O_{C_t}, expressed as:

$$\pi(a_t | W, O_t, O_{C_t}, S_t; \Theta)$$

This study conducts the VLN task in a zero-shot setting, where Θ is not trained using VLN-specific datasets but is instead derived from the language corpus used to train the LMs.

Dataset Translation

To convert the English R2R dataset to Arabic, we used the Groq API, specifically the Llama-3.2-90B-text-preview model. We then developed a simple prompt-based code to translate the English instructions, objects list, and observations in the R2R dataset to Arabic, ensuring that the translated dataset maintained the same format as the original English version for seamless compatibility with NavGPT. This alignment facilitates using both the original and translated datasets in comparative experiments.

Incorporating LMs

To explore different language models with NavGPT, we utilized LMs deployed on Azure for inference, enabling a comparative analysis. Following NavGPT's pipeline shown in Figure 1, we configured the SLMs (GPT-4o mini, Llama, Phi-3) and Jais 30B LLM, and tested them on the English and Arabic datasets to facilitate a direct comparison of their performances.

The selected SLMs were chosen for their diverse architecture sizes, multilingual capabilities, and unique approaches to processing input. GPT-4o mini, with its compact architecture, demonstrates efficiency in understanding and reasoning across languages by leveraging large-scale multilingual training (Vidhyashree 2024). Llama 3 8B excels in instruction-following tasks, combining a medium-sized model with strong contextual understanding (Dubey 2024). Phi-3 Medium 14B balances scalability and reasoning power, enabling nuanced task-specific performance (Abdin et al. 2024). Jais 30B, optimized for Arabic, enhances linguistic diversity by deeply integrating Arabic-specific datasets, ensuring accurate comprehension and generation (Sengupta et al. 2023). These models allow for analyzing the effects of multilingual input on reasoning, focusing on whether instructions are directly reasoned upon or internally translated.

Inference with NavGPT

The NavGPT framework integrates natural language instructions and visual observations for autonomous navigation. Instructions are processed alongside environmental data using Visual Foundation Models, which extract key features from the current viewpoint. A Prompt Manager formats this information into structured inputs for an LM, which reasons over the trajectory to decide the next action or stop. A history buffer tracks previous states to ensure consistent decision-making. This pipeline facilitates robust multimodal navigation with real-time reasoning capabilities tailored to user instructions.

We ran NavGPT in inference mode using the SLMs and Jais for a subset of the data. We assessed the configured LMs' effectiveness on English and Arabic instructions and evaluated their performance to see if language affects reasoning in navigation. We used 100 sample trajectories from the *val unseen* dataset in the R2R dataset. Each trajectory output includes:

- **Action Input**: The upcoming trajectory ID
- **Observation**: Textual descriptions of the environment at each location
- **Thought**: The robot's thoughts, reasoning, and planning about reaching the target location and identifying obstacles
- **Evaluation Metrics**: Action steps, total steps, path lengths, navigation error, oracle error, success rate (SR), oracle success rate (oracle SR), success weighted by path length (SPL), normalized dynamic time warping (nDTW), success weighted by dynamic time warping (SDTW), and coverage length score (CLS)

Through these steps, our methodology combines translation, inference, and evaluation, providing a structured approach to deploying NavGPT for Arabic-language navigation tasks. This approach ultimately allows us to measure the model's efficacy across languages and guide improvements for multilingual robotic navigation.

Experimental Setup

The experiments were conducted using a combination of local hardware and cloud-based APIs. The local setup included machines with NVIDIA Quadro 6000 GPUs, each with 24 GB of memory, primarily for dataset preparation and evaluation tasks. The models evaluated—GPT-4o mini (Vidhyashree 2024), Llama 3 8B (Dubey 2024), Phi-3 Medium 14B (Abdin et al. 2024), and Jais 30B (Sengupta et al. 2023)—were hosted on Azure's serverless platform and accessed via APIs in the same configuration. This setup ensured consistency in model performance while leveraging Azure's scalability. For dataset augmentation, the Groq API was used to generate Arabic translations of English instructions in the R2R dataset.

This study evaluates the zero-shot reasoning capabilities of pre-trained language models, focusing on their ability to handle navigation tasks in both English and Arabic. No training or fine-tuning was performed. Input instructions and navigation trajectories were fed directly to the models via APIs without modifications to the underlying model architecture. To ensure outputs were in the correct format, different prompts were used depending on the model, aligning responses with the required structure for evaluation. This consistent configuration allowed for a controlled and fair comparison across all models.

Dataset

The evaluation utilized the R2R dataset alongside its Arabic-augmented counterpart. Arabic translations were generated using the Groq API, maintaining alignment with the original English instructions. A total of 100 navigation trajectories were evaluated in each language. The same 100 trajectories from the English dataset were used in the augmented Arabic dataset to ensure consistency. This framework allowed for a direct comparison of language-specific model reasoning and navigation capabilities.

Evaluation Metrics

Quantitative assessment We compared the models with each other using the following standard evaluation metrics for VLN tasks (Anderson et al. 2018):

- **Trajectory Length (TL)**: the average distance traveled by the agent during navigation
- **Navigation Error (NE)**: the mean distance between the agent's final location and the target location
- **Success Rate (SR)**: measures the percentage of completed trajectories where the robot reaches its goal
- **Oracle Success Rate (OSR)**: evaluates whether the agent was on the right path even if it didn't stop at the exact target location
- **Success weighted by Path Length (SPL)**: considers the length of the trajectory relative to the shortest path

Quantitative results are shown in Table 1.

Qualitative Assessment We also conducted a qualitative assessment to examine the models' performance, focusing on their reasoning and decision-making processes. This evaluation highlights subjective observations that help identify weaknesses in the models' reasoning and planning capabilities. Specifically, we assessed the following aspects:

- **Reasoning**: Analyzed how effectively the models interpreted navigation instructions, integrated visual observations, and decomposed complex instructions into actionable sub-goals.
- **Spatial Awareness**: Evaluated the models' ability to comprehend their current environment, maintain navigation history, and use this information to make accurate decisions.

The qualitative evaluation was performed using both Arabic and English instructions, providing insights into how the input language affected the models' behavior and reasoning processes. Figure 3 shows an example of an agent successfully understanding the instruction and navigating to the desired area. Other failure cases are discussed in more details in the appendix section.

Results and Discussion

This study explored the performance of various language models in reasoning and understanding complex navigation

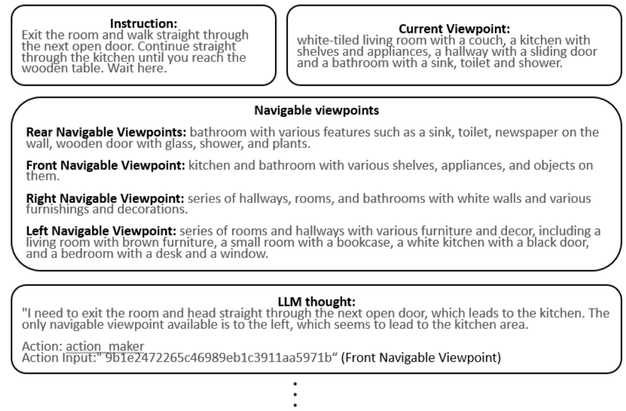

Figure 3: Successful navigation example

instructions in English and Arabic. The models tested included GPT-4o mini, Llama 3, Phi-3, and Jais, with significant variations in their ability to parse and execute instructions. Furthermore, the appendix provides textual examples of some of the models' outputs, including failed cases. Based on the models' performance, we categorized them into three groups: Working, Partially Working, and Not Working.

Table 1 presents the aggregated results of quantitative evaluation metrics. During inferencing, NavGPT outputs thoughts that demonstrate its reasoning process as it navigates the environment. We prompted the model to output its thoughts in Arabic whenever we were inferencing with the Arabic-translated dataset as input, creating a monolingual Arabic context. Only with GPT-4o mini did we mix input languages, combining Arabic datasets and English thoughts, to observe how this robust model performs.

We evaluated the models on 100 trajectories. However, some of the language models frequently ran out of context window when reasoning through more complex long instructions. This occurred primarily with smaller or less robust models, such as Phi and Jais, as reflected in their lower number of successful predictions out of 100 trajectories (Table 1).

NavGPT relies on structured prompts to perform optimally in navigation tasks. Specifically, the input should include a well-defined task description, such as goal location and intermediate waypoints. When these structured inputs are missing or incomplete, the model often struggles to generate accurate predictions, as shown in Figures 5 and 6. These errors typically happen when the model cannot fully reason over the provided instructions, resulting in its failure to output the necessary information in the required format for continued navigation.

- **Working:**
  1. **GPT-4o mini:** This model is trained on a large multilingual dataset, eliminating the need for explicit translations of non-English inputs. As a result, it successfully handles both English and Arabic datasets. Its performance metrics for monolingual English and Arabic

scenarios were comparable, achieving the highest values for Trajectory Length (TL), Success Rate (SR), Oracle Success Rate (OSR), and Success weighted by Path Length (SPL), while maintaining the lowest Navigation Error (NE) compared to other models, showcasing its robustness. However, in the mixed scenario of Arabic data with English reasoning, it experienced slightly higher navigation error, lower SR and SPL, and a marginally higher OSR (37.00) compared to the pure Arabic scenario (36.00). This discrepancy could be attributed to misalignment between the Arabic dataset and English reasoning. Nevertheless, GPT-4o mini outperformed the next best model, Phi-3 (SR=7), by approximately three times (SR=21), and five times better than Llama 3 (SR=4).

2. **Llama 3 8B:** Llama 3 exhibited reasonable reasoning and planning capabilities when processing both English and Arabic instructions. However, it fell short of GPT-4o mini's performance metrics, likely due to its smaller multilingual training dataset and less optimization for diverse linguistic tasks. Its SR (4 for English and 3.12 for Arabic) indicates limited success in executing goal-oriented tasks. Despite these challenges, Llama 3's ability to handle Arabic instructions suggests it holds promise for future development in multilingual reasoning. Even with its smaller size (8B), Llama 3 successfully completed nearly all 100 trajectories with both datasets, demonstrating its robust capabilities and large context window.

- **Partially Working (Phi-3 medium):**

  Phi-3 medium demonstrated competitive performance in processing English instructions but faced challenges due to a smaller number of successful predictions (41/100) and parsing issues with the *viewpoint ID*. These issues were likely caused by the model's strict format requirements for input alignment, highlighting its lack of robust natural language understanding. This limitation led to incorrect outputs, as detailed in the appendix.

  For Arabic tasks, Phi-3 failed entirely, which can be attributed to its non-multilingual nature and insufficient exposure to Arabic language data during training. Consequently, it was unable to process or generate meaningful outputs in Arabic. Moreover, the model only evaluated 18 out of 100 trajectories, revealing its limited robustness in handling complex tasks in Arabic.

- **Not Working (Jais):**

  Jais 30B, the only Arabic-centric LLM in this experiment, surprisingly exhibited poor reasoning capabilities in both Arabic and English, performing the worst across all critical metrics. Although it was expected to perform well due to its Arabic focus, Jais 30B struggled with reasoning in the context of navigation tasks. This poor performance may be attributed to its initial training by Core42, which was not specifically optimized for navigation-related tasks. However, despite these limitations, Jais 30B's large size allowed it to complete 82 out of 100 trajectories.

Model	Data	Succ.	TL	NE↓	SR↑	OSR↑	SPL↑
GPT-4o mini	Eng	100	17.6	6.98	21.0	46.0	13.0
	Ar	100	17.7	7.18	20.0	36.0	9.34
	Mixed	100	17.1	7.87	16.0	37.0	8.08
Phi-3 med	Eng	41	6.89	7.65	7.32	7.32	5.66
	Ar	18	2.36	8.51	0.00	0.00	0.00
Llama 3 8B	Eng	100	8.21	8.20	4.00	8.00	2.73
	Ar	96	7.54	8.34	3.12	5.21	1.33
Jais 30B	Eng	95	0.68	8.45	0.00	0.00	0.00
	Ar	82	0.78	8.35	0.00	0.00	0.00

Table 1: Quantitative Analysis of the LMs with English and Arabic Datasets

Limitations and Future Work

The limitations of the proposed work can be summarized as follows:

- **Lack of Visual Features:** The visual images are not directly processed using an image encoder; instead, an image-to-text descriptor is used, which results in an information loss. As a result, we only depend on the textual depiction of visual scenes for language models.
- **Zero-Shot Reasoning Ability:** The language models used in this study were not fine-tuned for the specific downstream task. Instead, we relied on their zero-shot reasoning and planning capabilities for unseen tasks, which fell short compared to the fine-tuned models (Zhou, Hong, and Wu 2023).
- **Object History Tracking:** The history module summarizes previous observations into a sentence, which may result in omitting some details from earlier observations.
- **Translated Dataset (Instructions, Observations, Objects List):** Machine translation, especially for linguistically complex languages like Arabic, is rarely perfect and can introduce errors or ambiguities. This can lead to information loss, altered context, or misrepresented semantics, which in turn affects the model's ability to generalize and make accurate predictions.

Future work can address these limitations by incorporating a dedicated vision encoder to directly process visual features, avoiding the information loss caused by text-only descriptions. Additionally, datasets can be translated to Arabic by human annotators to improve the quality and accuracy of instructions. Exploring state-of-the-art Arabic-centric models like SILMA (Team 2024) and ALLaM (Bari et al. 2024) offers potential for enhancing Arabic language support. Finally, training and fine-tuning models on task-specific training data instead of relying on zero-shot predictions is another promising direction.

Conclusion

This work explored the impact of language on VLN tasks by comparing multilingual SLMs with the Arabic-focused LLM, Jais, in processing navigation instructions in both English and Arabic. We augmented the R2R dataset with machine-translated Arabic instructions and evaluated performance within the NavGPT framework. The results revealed that the robustness of the model performing the reasoning mattered more than the language itself. GPT-4o mini demonstrated strong performance in both languages, outperforming Llama 3 8B by five times. However, other models, such as Phi-3 medium and Jais, struggled due to parsing issues and limited reasoning capabilities, with Jais scoring 0 SR in both languages. Phi-3's poor performance with Arabic was attributed to its non-multilingual nature. These findings emphasize the need for robust, multilingual models to improve autonomous systems in Arabic-speaking regions, where language-specific models may be insufficient.

Appendix

Arabic instruction and Arabic thoughts

An example of our translated dataset with Arabic instructions and Arabic thoughts is provided in Figure 4, which shows a successful example of planning and reasoning in Arabic language.

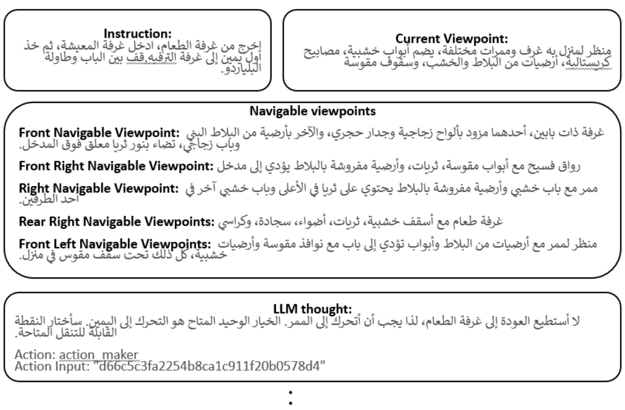

Figure 4: Example of Arabic instructions and Arabic thoughts

Parsing Error

An error occurs due to the model outputting the thought and action not in the expected format. This could be due to incorrect formatting, missing action output, or hallucinating a non-existing viewpoint as shown in Figure 5.

Jais Failing

Jais often fails in the format shown in Figure 6. It either indicates an inability to perform a navigation task or fails to comprehend the question. This behavior could stem from the lack of instruction-based fine-tuning for such tasks and the constraints imposed by the system prompt, which may limit its ability to plan navigation tasks.

References

Abdin, M.; et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

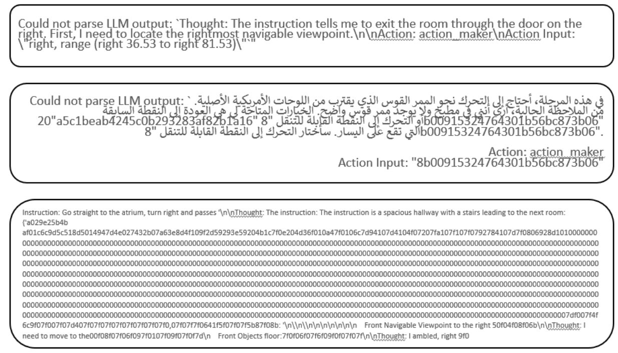

Figure 5: Example of parsing errors

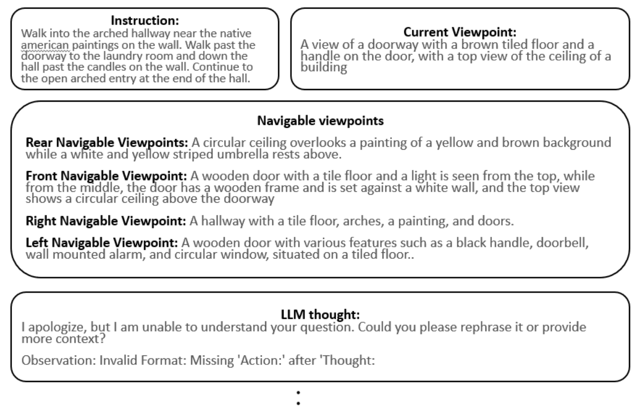

Figure 6: Example of Jais failing

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and van den Hengel, A. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. arXiv:1711.07280.

Bari, M. S.; Alnumay, Y.; Alzahrani, N. A.; Alotaibi, N. M.; Alyahya, H. A.; AlRashed, S.; Mirza, F. A.; Alsubaie, S. Z.; Alahmed, H. A.; Alabduljabbar, G.; Alkhathran, R.; Almushayqih, Y.; Alnajim, R.; Alsubaihi, S.; Mansour, M. A.; Alrubaian, M.; Alammari, A.; Alawami, Z.; Al-Thubaity, A.; Abdelali, A.; Kuriakose, J.; Abujabal, A.; Al-Twairesh, N.; Alowisheq, A.; and Khan, H. 2024. ALLaM: Large Language Models for Arabic and English. arXiv:2407.15390.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*.

Dubey, A. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Gu, J.; Stefani, E.; Wu, Q.; Thomason, J.; and Wang, X. 2022. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Lin-*

guistics (Volume 1: Long Papers). Association for Computational Linguistics.

Khalati, M.; Ali, T.; and Al-Romany, T. 2024. Artificial Intelligence Development and Challenges (Arabic Language as a Model).

Koto, F.; Li, H.; Shatnawi, S.; Doughman, J.; Sadallah, A. B.; Alraeesi, A.; Almubarak, K.; Alyafeai, Z.; Sengupta, N.; Shehata, S.; Habash, N.; Nakov, P.; and Baldwin, T. 2024. ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic. arXiv:2402.12840.

Ku, A.; Anderson, P.; Patel, R.; Ie, E.; and Baldridge, J. 2020. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. arXiv:2010.07954.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. arXiv:1908.02265.

Lu, Z.; Li, X.; Cai, D.; Yi, R.; Liu, F.; Zhang, X.; Lane, N. D.; and Xu, M. 2024. Small Language Models: Survey, Measurements, and Insights. arXiv:2409.15790.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Rahmanzadehgervi, P.; Bolton, L.; Taesiri, M. R.; and Nguyen, A. T. 2024. Vision language models are blind. arXiv:2407.06581.

Sengupta, N.; Sahu, S.; Jia, B.; Katipomu, S.; Li, H.; Koto, F.; Afzal, O.; Kamboj, S.; Pandit, O.; Pal, R.; Pradhan, L.; Muhammad Mujahid, Z.; Baali, M.; Aji, A.; Liu, Z.; Hock, A.; Feldman, A.; Lee, J.; Jackson, A.; and Xing, E. 2023. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models.

Tan, H.; Yu, L.; and Bansal, M. 2019. Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. arXiv:1904.04195.

Team, S. 2024. Silma.

Vargas-Munoz, J. E.; Srivastava, S.; Tuia, D.; and Falcao, A. X. 2021. OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9(1): 184–199.

Vidhyashree, A. 2024. GPT-4O Mini: The New Lightweight Version of GPT-4. [Accessed: Nov. 19, 2024].

Zhang, Y.; Ma, Z.; Li, J.; Qiao, Y.; Wang, Z.; Chai, J.; Wu, Q.; Bansal, M.; and Kordjamshidi, P. 2024. Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models. arXiv:2407.07035.

Zhou, G.; Hong, Y.; Wang, Z.; Wang, X. E.; and Wu, Q. 2024. NavGPT-2: Unleashing Navigational Reasoning Capability for Large Vision-Language Models. arXiv:2407.12366.

Zhou, G.; Hong, Y.; and Wu, Q. 2023. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. arXiv:2305.16986.