Arabic Dialect Detection using LLMs

Senior

Progress Report 1

Student Name	UID
EmadEddin Abdulsalam Al-Chmri	*U22105598*
Moheeb Suliman Musa Suliman	*U22105916*
Isa Al-Khanous	*U22200681*
Mohammad Ra'ed Mohammad Hardan	*U22105630*

Computer Science

Supervisor: Dr. Ashraf Mohamed Al-Nagar

College of Computing and Informatics

University of Sharjah

October 2025

Contributions

Student Name	Role	Contributions
EmadEddin Abdulsalam Al-Chmri	Principal Trial Investigator, Prompt Engineer, Data Analyst	Conducted and managed trials, modified prompts, produced trial analytics and charts.
Moheeb Suliman Musa Suliman	Theory Researcher, Trial Co-Investigator, Scope Planner	Researched new approaches, conducted trials, managed the scope of the project.
Isa Al-Khanous	Model Engineer, Programmer, Resource Manager	Fine tuned models, coded python scripts, provided resources for the team.
Mohammad Ra'ed Mohammad Hardan	Model Researcher, Trial Co-Investigator, Execution Planner	Researched new models, conducted trials, planned applications of proposed approaches.

Abstract

As we tackle a new stage of our project, it is of great importance to reflect on what was done previously and expand upon it, build it up to something worthy of proper display and application. In the past weeks, we aimed to do just that; progressing our testing through novel methodologies in order to achieve the goal described by our Junior Presentation.

Though it was relatively brief, the past weeks have led us to discover new dimensions to our project that we have not considered beforehand, of both the beneficial and daunting kinds alike. We will strive to accomplish the project to our best capabilities, and this report will detail how we endeavored to do so.

Research

In collaboration with our Supervisor, we spent our meetings discussing innovative directions to our project, some more promising than others.

The first initiative to be done was to conduct research regarding new Large Language Models (LLMs) released during the period between the end of our Junior Presentation and the start of our Senior Project. We came across a handful of LLMs that seemed capable of achieving our goals, such as: the UAE's Falcon[1] developed by the Technology Innovation Institute; the KSA's ALLaM[2] developed by the Saudi Data and AI Authority; and the upcoming HUMAIN CHAT[3], another LLM developed by Saudi Arabia.

In addition to newcomers, we analyzed the prospect of version updates of previous contenders such as Open AI's ChatGPT[4], X's Grok[5], and Alibaba Cloud's Qwen[6]. Due to our Junior results indicating the suitability of these LLMs for our purposes, specifically Grok's staggering degree of accuracy, they held high hopes regarding any improvements exhibited by them.

In order to expand our set of viable LLMs, we came to the conclusion that our methods of testing needed some adjustments, primarily to unearth any possibility to increase the accuracy of our top performing LLMs. One such adjustment was to conduct Mult-Label testing, wherein we allowed each LLM to submit up to three predictions about the sentence's region of origin. This would help detect if any of the LLMs' second or third guesses would be correct due to some sentences' ambiguous nature.

Another experiment to conduct was by way of fine tuning some LLMs, to increase their accuracy. This required a relatively more dedicated approach that was both tedious and a logistical hurdle we had to overcome to gain any valuable results. The learning process and subsequent trials of using the popular LLM hosting website HuggingFace were extremely valuable as we delved into the core of what made LLMs work on a base level.

We were also motivated by our Supervisor's own discoveries. He informed us about the possibility of creating a Multi-Agent system in which a sentence submitted to a frontend agent would be sent to each of our top

performing LLMs, and they would vote about the predicted origins of the sentence, then parse in a reply that would be displayed by the frontend agent. This required much more research and experimentation, but might prove to be the best application of our proposed system.

Challenges

Though we have spent our efforts during our Junior project into making a suitable and quick workflow, it was clear that the challenges we faced in this Senior period came from unexpected complications, some entirely out of our control.

One such inevitability was that more and more LLMs would be produced and released to the public, some even updating during the testing process. Because of our commitment to thorough experimentation, we had to simultaneously keep up with all the advancements happening in the LLM scene whilst maintaining our pace with other commitments. Thankfully, our Supervisor helped us stay motivated throughout.

Out of all challenges faced during the beginning of our Senior project efforts, the least to surprise us is Arabic LLMs' incompetency. Though developed by Arabic institutes, it has come to be expected they will fare the worst when compared to foreign-based LLMs. This is best demonstrated by Jais, as it did not accept any trial input due to "detection of indecent words" in the prompt, which did not exist in the dataset. Additionally, the Falcon-Arabic model's results were reduced to repeated predictions and incoherent rambling, which is a shame considering Falcon-H1-43B-Instruct's above-average performance.

An additional hurdle was our prompt engineering. At the tail end of our Junior journey, we seemed to develop a "perfect" prompt that was extremely extensive and provided accurate results. However, due to its admittedly absurd length, it led to some LLMs not operating as best they could as the

amount of tokens was not something they can handle in conjunction with the dataset. This made us fall back on previous concise prompts that served us well in the end.

The most challenging aspect of our dive into more hands-on LLM experimentation was with fine tuning and the use of HuggingFace, a platform that hosts models that can be interfaced with directly through python code. The learning curve for using such a service was tumultuous to say the least, as none of us had any experience with such a service. Eventually, we tried using some of the LLMs like ALLaM-7b and LLama-3[7]. We, however, could not gain any valuable data due to the limited tokens provided by the service.

A daunting realization we had to face was the sheer scope of the project, now that we are approaching ever closer to its end. We were faced with a plethora of new options in order to implement our final product, including multi-agent systems and programming our own special agent for parsing requests and responses. We would have to rise to the challenge in order to achieve something so out of our comfort zone such as this.

Progress

Our progress during the weeks prior was scattered, with a significant amount of results coming in bursts as we uncovered more and more approaches and decided on which to prioritize.

The first on the agenda was testing the newcomer LLMs. Detailed below is a comprehensive list of the models tested:

Index	Model
1	Falcon (TII) - H1 - 43B - Instruct
2	ALLaM - 7b
3	Jais[8]
4	LLama - 3
5	K2 Think[9]
6	GPT-OSS-20b

The second set of experiments is Multi-Label testing, in which we instruct the LLM to provide its top three predictions instead of top one. This is done in order to prove that some LLMs were indecisive when choosing a single label and, if given multiple chances, will increase their accuracy. If any of the three predicted labels were true, it would be considered an accurate classification.

The models used for this procedure were ones that achieved higher than fifty-percent during zero shot testing done beforehand, to observe how our best candidates can improve. The changes in their accuracy will be noted in the averages table below the multi-label results charts.

Index	Model
1	Grok
2	Qwen
3	DeepSeek[10]

Our approach remained the same throughout:

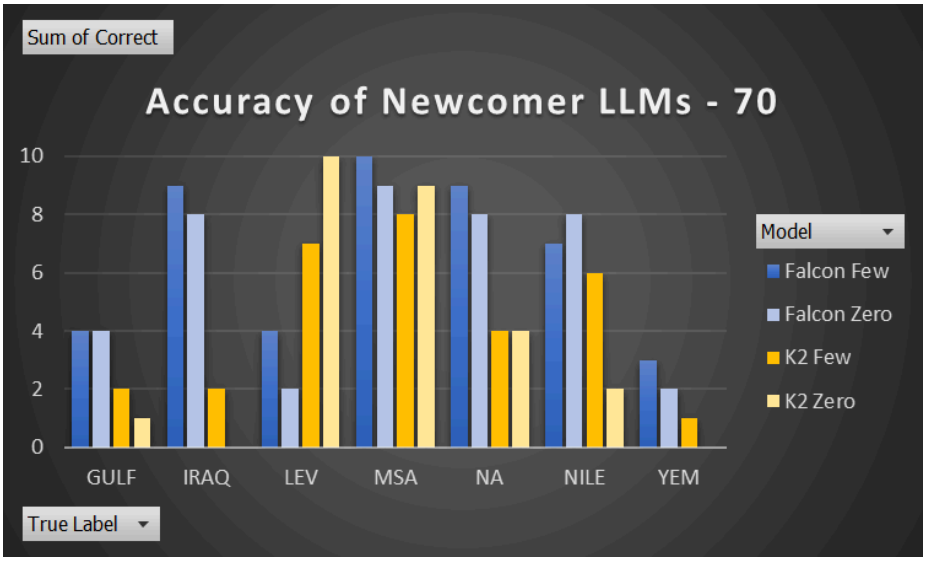
1. Enter the engineered prompt.
2. Add the .xlsx file or sentences as input.
    a. Trial 1: 70 sentences from the 210.
    b. Trial 2: 140 sentences from the 210.
    c. Trial 3: All 210 at once.
3. Transfer the results into an Excel sheet.
4. Compare the results with the true labels.
5. Create pivot tables concerning various elements.
    a. The model's predicted label variety per category, if applicable.
    b. The model's accuracy per category.

This standardized all of our results to a comparable degree, which was essential in determining the best performers. As was expected, some of them provided us with unworkable or disappointing results, but we understood that we had to conduct these failing tests for the sake of thoroughness, and so we can say we left no stone unturned.

We would like to state that, due to the described challenges above, only Falcon and K2 Think passed the trials with no complications. As such, we will only display their results in this report.

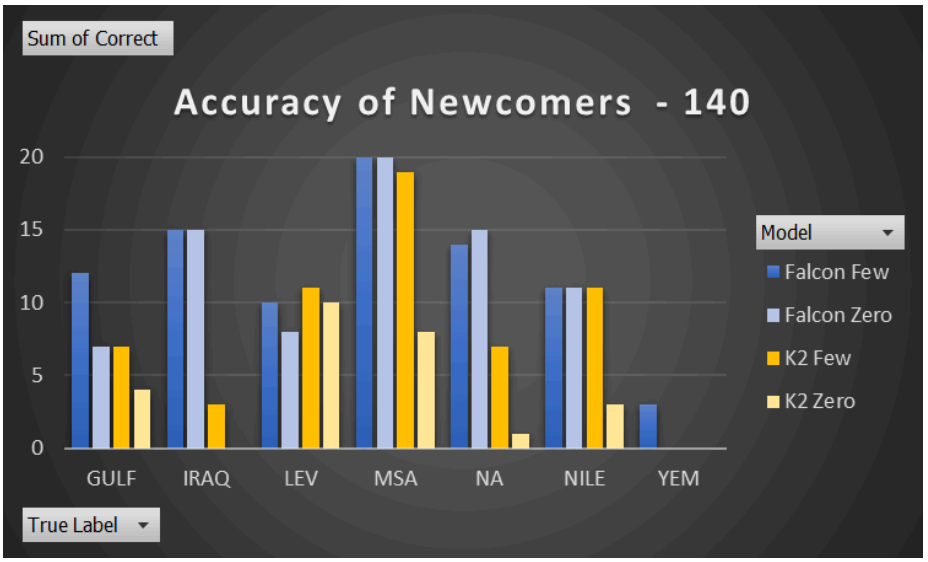
Newcomer Trials

Figure 1.1: Trial 1 - Accuracy of Newcomers per Category



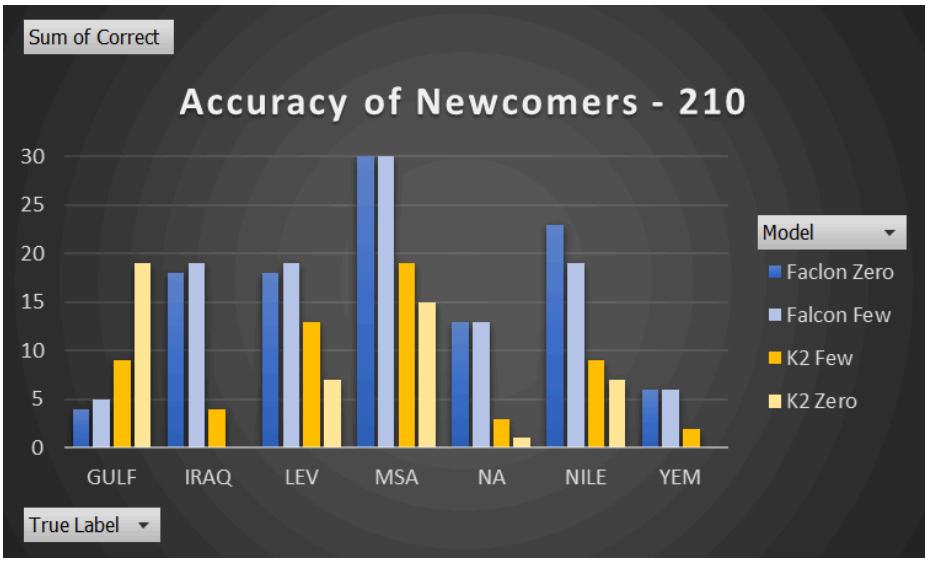
Model	Falcon		K2 Think	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	58.57%	65.71%	37.14%	42.86%

Figure 1.2: Trial 2 - Accuracy of Newcomers per Category



Model	Falcon		K2 Think	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	54.29%	60.71%	18.57%	41.43%

Figure 1.3: Trial 3 - Accuracy of Newcomers per Category



Model	Falcon		K2 Think	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	53.33%	52.86%	23.33%	28.10%

Averages:

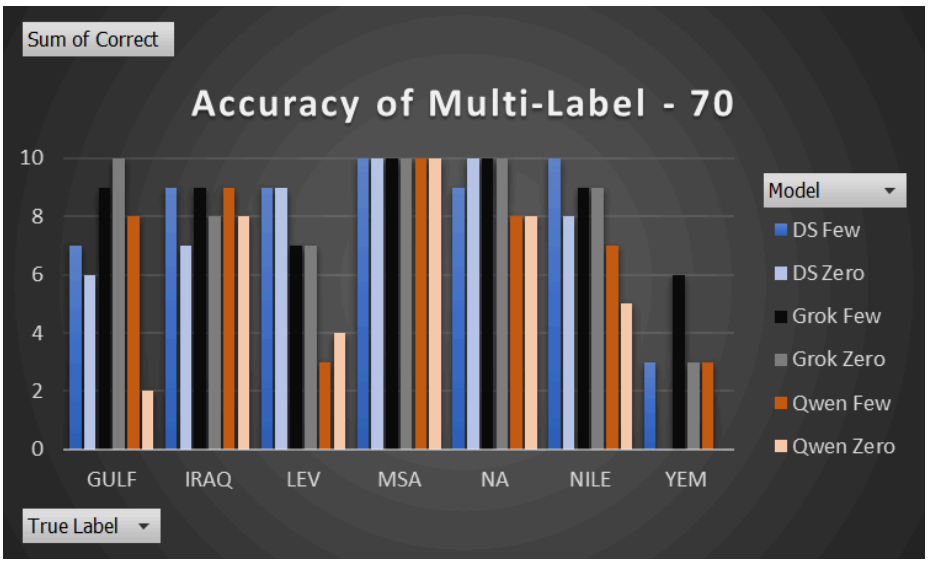
Model	Falcon		K2 Think	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	55.40%	59.76%	26.35%	37.46%

Although Falcon-H1-43B-Instruct showed promise, the gulf between it and the best performers like Grok and DeepSeek leaves it in contention against Qwen, which showed similar yet slightly better results. It is by far the most successful Arabic LLM we have ever tested.

In brief, K2 Think is the typical Arabic LLM; low accuracy and difficult to work with. It is remarkably similar to Fanar, our lowest performer, and like it, K2 will likely never be considered for our future endeavours.

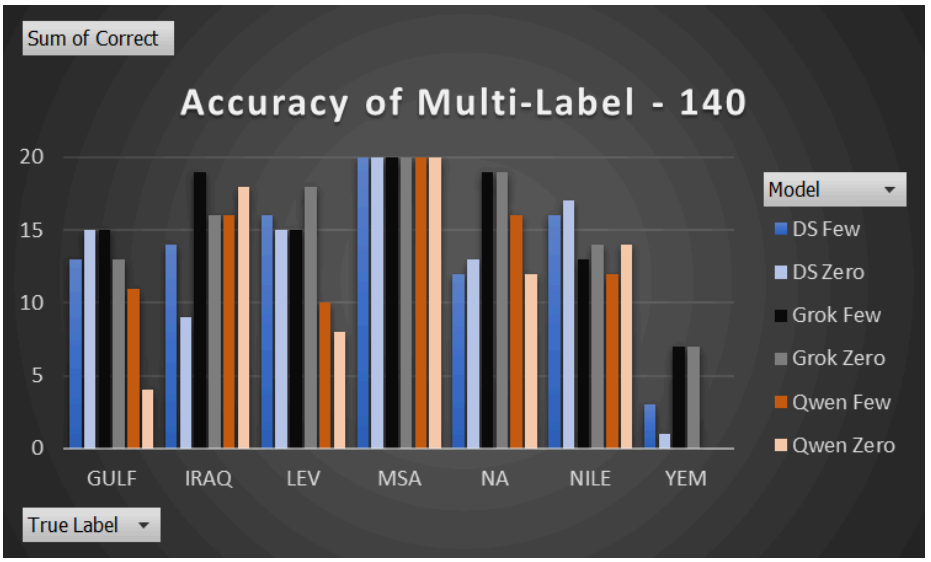
Multi-Label Trials

Figure 2.1: Trial 1 - Accuracy of Multi-Label per Category



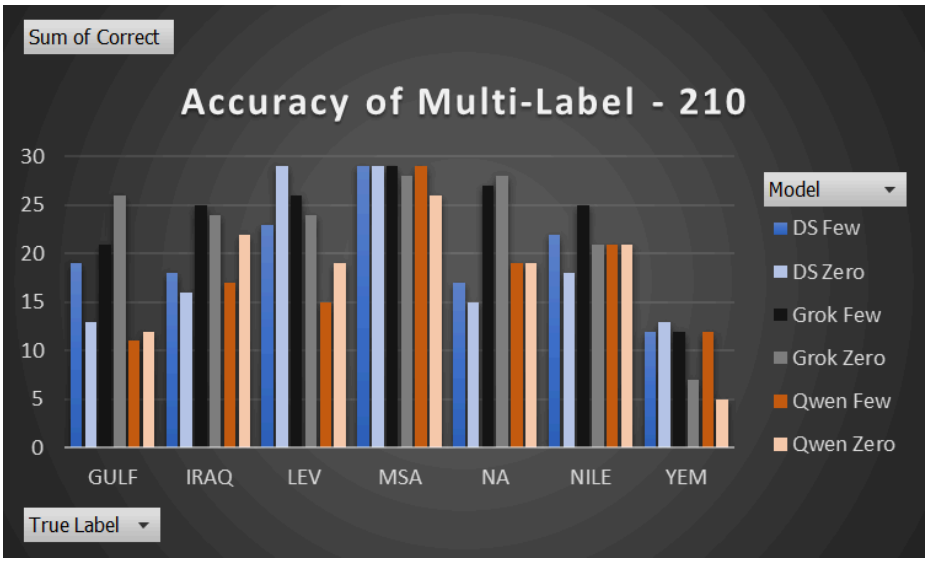
Model	DeepSeek		Grok		Qwen	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	71.43%	80.00%	81.43%	82.86%	52.86%	64.29%

Figure 2.2: Trial 2 - Accuracy of Multi-Label per Category



Model	DeepSeek		Grok		Qwen	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	64.29%	67.14%	76.43%	77.14%	54.29%	60.71%

Figure 2.3: Trial 3 - Accuracy of Multi-Label per Category



Model	DeepSeek		Grok		Qwen	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	63.33%	66.67%	75.24%	78.57%	59.05%	59.05%

Averages:

Model	DeepSeek		Grok		Qwen	
Type	Zero Shot	Few Shot	Zero Shot	Few Shot	Zero Shot	Few Shot
Accuracy	66.35%	71.27%	77.70%	79.52%	55.40%	61.35%
Change	+8.89%	+4.60%	+3.57%	+0.63%	-1.27%	-1.03%

When investigating Multi-Label trials, we were expecting a stark increase, which we have definitely achieved, though not to the extent we imagined. Statistically, having three guesses would provide it with greater accuracy, but almost all models opted not to utilize all chances, only rarely doing so with particularly tricky sentences. In broad strokes, few shot trials gained less from multi-labeling when compared to zero shot.

DeepSeek gained a sizable increase, as it was the model that took advantage of multi-labeling the most. This might imply hidden potential left to discover with DeepSeek, specifically with more developed prompts.

Grok's increases were marginal, yet it remained the best performing LLM by far, with an almost six-percent difference between its zero shot result and the highest few shot result.

Curiously, Qwen suffered from a decrease of around one-percent. Additionally, its zero shot trial results improved from the first to the last, which is unprecedented. This may require further testing to ascertain the cause of such strange behaviour, as Qwen is one of our top performers and has high priority.

Reflections

This stage of the project has proven to be both challenging and eye-opening. We entered the Senior phase believing our prior experience from the Junior project had prepared us for all foreseeable difficulties. However, the complexity of modern Large Language Models and their unpredictable performance with Arabic text forced us to adapt rapidly and think critically.

We discovered that success in LLM-based research is not limited to accuracy metrics. It also requires creativity in prompt design, flexibility in testing methodologies, and continuous learning as new models and techniques emerge. The process of fine-tuning and experimenting with HuggingFace, though initially intimidating, became an invaluable opportunity to understand model behavior at a deeper level.

Moreover, the collaboration between our team members and Supervisor was essential in maintaining direction and motivation. Dr. Ashraf's guidance led us to explore novel ideas such as the Multi-Agent voting system, which opened new possibilities for practical application. However, it was a shame that our schedules did not align as well as during our Project's Junior stage, which caused a few setbacks and confusion at first.

Finally, this period reinforced our belief that Arabic NLP is still a developing frontier with immense potential. While Arabic-native models underperformed, this motivated us to explore how hybrid or multi-model systems might compensate for their weaknesses and push accuracy beyond current limits.

Milestones Achieved

1- Comprehensive LLM Re-evaluation:

- Successfully identified and tested a collection of recently published LLMs (Falcon, ALLaM, Jais, LLama-3, K2 Think) in addition to the updated versions of our prior LLMs (DeepSeek, ChatGPT, Grok, and Qwen).

2- Innovative Multi-Label Testing:

- Designed and implemented an advanced Multi-Label testing procedure for top-performing models (Grok, Qwen, DeepSeek), resulting in a measurable increase in accuracy for Grok and DeepSeek.

3- Project Scoping and Direction:

- Engaged in productive discussions with our supervisor to explore and define future project directions, culminating in the proposal of a Multi-Agent system as the potential final product.

4- Technical Skill Development:

- Gained initial practical experience with the HuggingFace platform and the concepts of model fine-tuning, building foundational technical capacity within the team.

Future Milestones

1- Finalize Model Selection and Architecture Design:

- To choose the top three to five LLMs for system integration, we will compile our testing data. At the same time, we will complete the multi-agent system's technical architecture, including the protocols for data flow, interaction, and voting between the frontend agent and the LLM voter agents.

2- Implement the Frontend Agent and User Interface:

- We will focus on developing the user-facing component of the system, which involves designing and developing an intuitive web-based interface for Arabic sentence input. This work also includes building the frontend agent logic to manage user requests, facilitate communication with the backend multi-agent system, and clearly present the final dialect prediction along with potential consensus details.

3-End-to-End System Integration and Rigorous Testing:

- Following a thorough testing phase to confirm its functionality, we will combine the frontend agent, the LLM voter agents, and the consensus engine into a single, cohesive solution. This will entail confirming that the accuracy of the multi-agent system is on par with or better than the best individual models, testing it with long texts, ambiguous words, and unexpected formats to

ensure its resilience, and getting user input to confirm the interface's usability and clarity.

4- Preparation of Final Deliverables:

- We will compile all project outcomes, including the final integrated system, source code, a comprehensive project report detailing our methodology and findings, and a presentation, for the project's conclusion.

References

1. Falcon-H1-43B-Instruct and Falcon-arabic, available on: https://chat.falconllm.tii.ae/
2. ALLaM-7b-Instruct-Preview, available on: https://huggingface.co/humain-ai/ALLaM-7B-Instruct-preview
3. HUMAIN CHAT, will be released soon on: https://chat.humain.ai/
4. ChatGPT and its models, available on: https://chatgpt.com
5. Grok and its models, available on: https://grok.com
6. Qwen3-Max, available on: https://chat.qwen.ai/
7. Llama-3, available on: https://www.llama.com/models/llama-3/
8. Jais 30B and Jais 70B, available on: https://jais.inceptionai.ai
9. K2 Think, available on: https://www.k2think.ai/
10. DeepSeek and its models, available on: https://chat.deepseek.com/