Dinh-Tu Nguyen
Faculty of Mechanical Engineering
Can Tho University of Technology
Vietnam

Phuoc-Loi Tran
Faculty of Mechanical Engineering
Can Tho University of Technology
Vietnam

Chi-Ngon Nguyen
Faculty of Automation Engineering
Can Tho University
Vietnam

Huu-Phat Tran
Department of Mechanical Engineering
National Central University
Taiwan

Dinh-PhucTran
Graduate Institute of Manufacturing
Technology, National Taipei University
of Technology
Taiwan

LLM and NLP-based Drone Control: A Distributed Edge Computing Architecture for Vietnamese

This study proposes an edge-based Vietnamese voice-control system for unmanned aerial vehicles. The system integrates a large language model, natural language processing modules, and real drone actuation to improve the reliability of naturallanguagebased drone control. The main contributionisan embedded mechatronic control framework experimentally validated on a commercial drone platform. This framework combines a multi-layered prompt structure with a preprocessing mechanism to reduce repetitive-word, invalid command generation, and redundant textual output. The proposed system uses a distributed edge architecture. An NVIDIA Jetson Orin NX performs Llama 3.1:8b inference, while a Raspberryhandles PhoWhisper-based speech recognition and direct DJI Tello control. Under noise-controlled conditions, the system achieved 100% accuracy for single, compound, and inferential command categories. The average response time was 2–3 s for single commands and 8.45 s for a representative complex compound command.

Keywords: *Large language models, unmanned aerial vehicles, drones, edge computing, natural language processing, application programming interface, multi-threaded architecture.*

1. INTRODUCTION

Unmanned aerial vehicles (UAV) currently play an essential role across multiple domains, including tech–nical education, security surveillance, disaster rescue, logistics, infrastructure inspection, environ–mental monitoring, and precision agriculture [1-9]. Owing to their mobility and the integration of high-performance microprocessors, drones serve as promising platforms for research in autonomous systems and intelligent interaction [10-13]. However, UAV operation remains challenging for most users, necessitating complex coordination skills, structured training, and extended familiarization periods. Consequently, traditional manual control methods have demonstrated notable limitations in flexibility and accessibility for non-expert users [14, 15].

To simplify interaction interfaces, large language models (LLM) and natural language processing (NLP) are being integrated into control systems to enable voice-based command execution [16, 17]. This inte–gration allows the system to analyse input, comprehend context, and determine communicative intent, thereby generating natural responses or executing physical control commands. Nevertheless, deploying LLM in practical control systems remains constrained by issues concerning reliability and logical reasoning capabilities. A primary challenge is erroneous function calling, which stems from the inherent diversity and ambiguity of natural language [18-20]. Furthermore, current LLM frequently struggle to process complex, multi-step com–mand sequences, potentially resulting in safety const–raint violations or incorrect action sequencing[21].

To enhance the accuracy of voice-controlled UAV, Lytovchenko proposed a system integrating the Whisper speech recognition model with a sub-band keyword spotting module and weighted finite-state transducer grammatical constraints [22]. This solution enables the direct, real-time conversion of natural language com–mands into radio controlchannel control signals while minimizing the misrecognition of numerical parameters and directional markers. Experimental results indicate that the integrated model achieved an accuracy of 90.81% and reduced the word error rateto 8.2%, demonstrating reliability and practical applicability.

To address the planning limitations of LLM in complex and domain-specific robotic tasks, Singh et al. proposed theLLM-based robotic system planning fra–mework, which integrates a strictly structured semantic system. By coupling this semantic system with a hie–rarchical chain-of-thought technique and dataset fine-tuning strategies, the model accurately captures the operational context of the device. Experimental results within retail workflows indicate this approach directly improves the accuracy, robustness and throughput of the robotic system during task execution [23].

Recently, Alkasim et al. proposed a voice-controlled UAV system integrating automatic speech recognition (ASR) with LLM to optimize command execution. Experiments conducted in a robot operating system simulation evaluated ASR models (notably Whisper) and LLM (e.g., GPT-4o-mini, T5-Large, BART) for mapping transcribed text to micro air vehicle link control messages. Results demonstrate that Whisper

Received: March 2026, Accepted: May 2026
Correspondence to: Dinh-Phuc Tran
Graduate Institute of ManufacturingTechnology,
National Taipei University of Technology, Taiwan
E-mail: ndtu@ctuet.edu.vn
doi: 10.5937/fme2603402N

Base is optimal for speech recognition, while BART-Large-MNLI achieves the highest zero-shot command translation accuracy [24].

Liang et al. proposed the ASR system to address the limitations of existing research, which primarily applies LLM to small-scale aerial vehicles for single-task operations. To accommodate the complex operational requirements of medium- and long-range UAV, this framework employs an LLM as a parser to translate natural language commands into points of interest. Detailed coordinate computation and kinematic execution are subsequently delegated entirely to traditional trajectory planning algorithms to ensure precision & physical safety. In addition to demonstrating efficacy in multi-UAV coordination scenarios, this study establishes a 5-level automation roadmap for integrating LLM into autonomous systems. [25].

This paper presents a novel integrated solution to address the aforementioned challenges. Specifically, this study proposes an integrated prompt framework to enhance the reasoning accuracy of LLM for controlling a DJI Tello drone via natural language. The system is based on a distributed edge computing architecture utilizing a Jetson Orin NX and a Raspberry Pi 5 to ensure stable processing performance. From an engineering perspective, the main contribution of this work is the development and experimental validation of an intelligent embedded mechatronic control architecture that converts Vietnamese natural-language commands into physical UAV motion. Unlike studies limited to text-based command interpretation or simulation, the proposed system was implemented on real hardware and evaluated through actual drone operation. This capa–bility provides a practical basis for more accessible human–UAV interaction in education, laboratory train–ing, indoor inspection, and other low-risk UAV opera–tion scenarios, particularly for non-expert Vietnamese users.

The main contributions of this research are as follows: (i) Improving the accurate execution of com–plex composite command sequences: The system enab–les the precise parsing and execution of multiple consecutive actions within a single query, enhancing the drone's operational flexibility in practical scenarios; (ii) Developing a multi-tier integrated prompt framework: The structure incorporates operational guidelines, a skill API directory, system constraints, and illustrative examples to guide accurate model reasoning and mini–mize function-calling errors; (iii) Optimizing the distributed hardware architecture: Processing tasks are allocated between a Raspberry Pi 5 (NLP, audio recording, and drone control) and a Jetson Orin NX (LLM inference and function calling) via an Ethernet connection. This approach maintains low latency and mitigates wireless signal interference during system operation.

The remainder of this paper is organized as follows: Section 2 presents the methodology, including system architecture design, hardware selection, and software development. Section 3 describes the experimental setup and evaluation methods based on three primary test command categories: single, composite, and complex commands. Section 4 details the experimental results, evaluating the proposed system's performance using accuracy and average response time metrics. Finally, we discuss the technical limitations of the system and outline potential future research directions for this case study.

2. MATERIAL AND METHODS

The methodology of this study is based on a modular edge-computing principle, in which the voice-control pipeline is divided into four functional layers: speech acquisition, speech-to-text conversion, natural language command interpretation, and UAV command execution. This design separates perception, reasoning, and actuation tasks to improve system stability and facilitate independent optimization of each module. The proposed framework is not limited to a specific UAV or computing device; rather, the selected hardware components are used as an implementation case to validate the feasibility of the architecture. In this structure, only predefined and safety-constrained control functions are executable, while ambiguous or unsupported commands are rejected or mapped to a safe fallback state.

2.1 Distributed hardware platform

To implement and validate the proposed modular edge-computing framework, the DJI Tello drone was selected as the experimental UAV platform (Figure 1). The DJI Tello was selected because it provides a suitable balance between safety, programmability, cost, and compatibility with indoor edge-computing experiments. Its compact and lightweight structure makes it appropriate for controlled laboratory testing, where safety and repeatability are critical. More importantly, the Tello supports command transmission through a Wi-Fi-based software development kit interface, allowing high-level control commands generated by the LLM function-calling module to be directly mapped to physical drone actions without requiring low-level flight-controller development. This makes it suitable for validatingthe proposed Vietnamese voice-based command interpretation framework. In addition, the platform has been widely adopted in research and education as an affordable UAV testbed, which improves the reproducibility of the proposed system. This lightweight UAV platform is equipped with an 1100 mAh lithium-polymer battery and provides a maximum flight time of approximately 13 minutes, and supports a maximum control range of up to 100 m under ideal conditions[26]. Control commands from the Raspberry Pi are transmitted to the drone through its Wi-Fi interface.

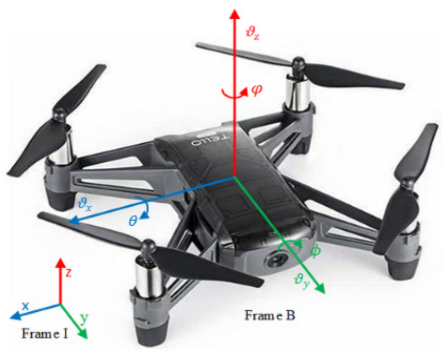

Figure 1. The DJI Tello UAV used as implementation [27]

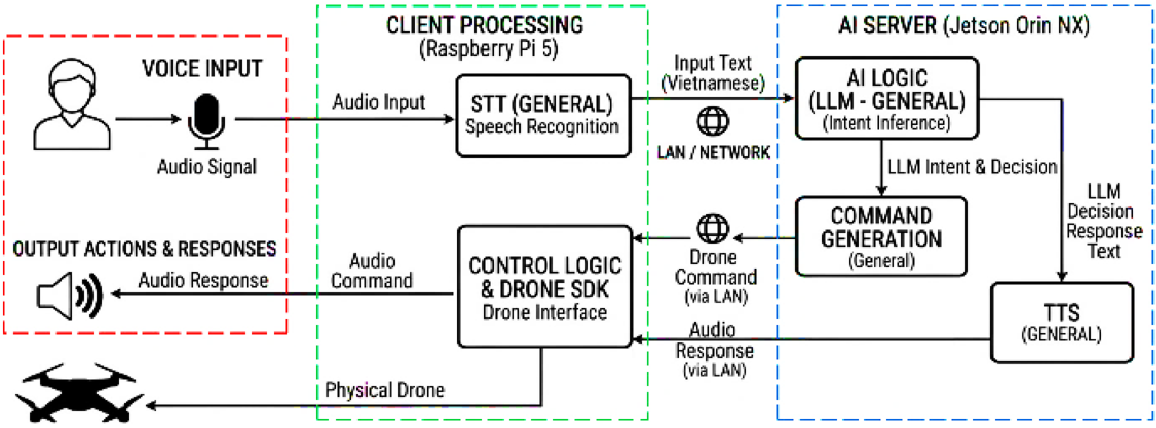

Figure 2. Architecture of the edge-based LLM system for natural language drone control

To address the resource constraints of an edge computing system and ensure low-latency large langu–age model (LLM) inference, the system employs a distributed computing architecture comprising two primary processing nodes: an NVIDIA Jetson Orin NX and a Raspberry Pi 5. This configuration facilitates the decoupling of input/output (I/O) intensiveNLP tasks from graphics processing unit (GPU) intensive deep learning inference tasks. The following Figure 2 illus–trates the overall design of this architecture.The NVIDIA Jetson Orin NX serves as the central inference server. Equipped with a 1024-core NVIDIA Ampere GPU and 32 Tensor cores. Featuring an 8-core ARM Cortex-A78AE central processing unit (CPU) and a memory bandwidth of 102.4 GB/s, the device meets the memory and processing speed requirements for executing language models such as Llama 3.1:8b at the edge devices.

The Raspberry functions as the direct user interface and peripheral management unit. Its primary tasks include: (i) acquiring and preprocessing audio signals via an omnidirectional microphone; (ii) executing the PhoWhisperspeech recognition model for speech-to-text conversion; and (iii) directly transmitting control commands to the drone via a wireless network interface. Additionally, a touchscreen enables users to monitor system status and interact visually through a graphical user interface (GUI).The two processing nodes are connected via a local area network (LAN) using gigabit ethernet interfaces. This wired connection ensures low transmission latency and data integrity during transfers between the edgeprocessing node (client) and the central processing node (server).

2.2 Multithreadedsystemintegration

The overall system is configured with a split-network architecture (Figure 3). The Raspberry Pi 5 operates as a LAN client (connected to the Jetson) while concurrently maintaining a Wi-Fi link with the drone. Allocating distinct internet protocol (IP) address ranges prevents packet collisions, ensuring uninterrupted control data flows during operation. This integration yields a highly reliable system capable of complex multitasking within a resource-constrained embedded environment.

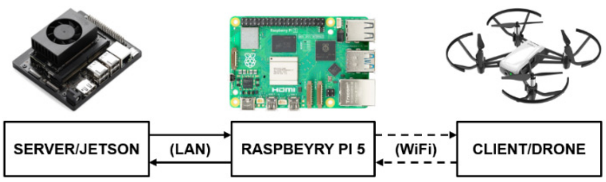

Figure 3. LAN and Wi-Fi split-network configuration

The Raspberry Pi 5 system employs a multithreaded architecture, enabling the parallel execution of diverse tasks, including audio I/O processing, speech recog–nition computation, and user interface management. Python’s built-in threading module is utilized to par–tition the operational workflow into three independent processing threads, as illustrated in Figure 4 below.

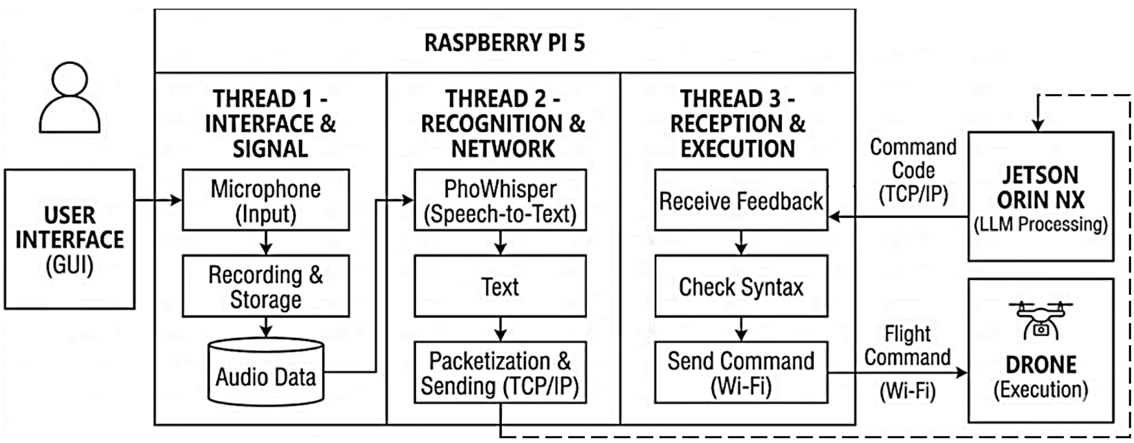

Figure 4. Diagram of the task threading mechanism on the Raspberry Pi 5 within the control system

As shown in Figure 4:

- Thread 1 - this thread maintains the GUIand manages virtual button logic. Upon command activation, it interfaces directly with the microphone to record and store audio data. This thread isolation ensures the GUI remains responsive and uninterrupted by background computational processes;
- Thread 2 - upon audio file creation, this thread executes the PhoWhisper model for speech-to-text conversion. The result is subsequently packaged and transmitted to the Jetson Orin NX via theIP protocol over the LAN;
- Thread 3 - this thread operates in a listening state to receive responses from theJetson Orin NX. Upon receipt of a control command, the system performs syntax validation and transmits the instruction to the drone via the Wi-Fi connection.

2.3 Speechprocessingmodule

The speech processing module serves as the interactive interface between the user and the control system. Its processing pipeline comprises two stages: Vietnamese ASR and text-to-speech synthesis for system feedback.

Audio data is acquired via a universal serial bus (USB)connected omnidirectional microphone. The duration of each recording session is controlled through a virtual button on the GUI. Signal acquisition occurs only while the button is held and terminates imme–diately upon its release. Continuous audio signals $x(t)$ are discretised into a signal sequence$x[n] = x(nT_s)$ and stored in uncompressed waveform audio fileformat to preserve spectral characteristics. Sampling frequency$f_s = 1/T_s$ is configured at a fixed 16 kHz, compatible with the input requirements of the PhoWhisper architecture[28].

PhoWhisper is a Vietnamese-optimized variant ba–sed on the OpenAI Whisper architecture, capable of effectively processing specific phonetic features and diacritics[29]. This study utilizes a PhoWhisper variant with approximately 74 million parameters. This selec–tion represents a balance between recognition accuracy and processing performance within the constrained resources of the Raspberry Pi 5[30, 31].

Through Hugging Face's Transformers library[32], signal $x[n]$ transformed into a Mel-spectrogram feature sequence$X$. The model then proceeds to decode the information to find the vocabulary sequence $\hat{W}$ the best solution is based on maximizing the conditional probability[33]:

$$\hat{W} = \arg\max_{W} P\left(W \mid X;\theta_{ASR}\right) \tag{1}$$

The system configures the θ_{TTS} parameter using the vais1000 voice package in open neural network exc–hange format, which is recognized as one of the most natural and stable open-source Vietnamese synthesis solutions currently available[34]. Furthermore, due to its lightweight architecture and rapid response time, it enables audio output via a USB audio adapter and external speakers without significant CPU overhead on the embedded system[35].

2.4 LLM-driven command mapping

The core of the natural language-based control system lies in the precise translation of relatively ambiguous and abstract textual requirements into specific technical execution commands[36]. This study employs Meta's Llama 3.1:8b model[37], deployed via the Ollama fra–mework to optimize performance on embedded devices[38]. The selection of this model is based on experimental comparisons with other model series, such as Qwen3:4b and Qwen3:8b. The following table sum–marizes the experimental results and evaluations from these tests.

As shown in the comparative results in the table above, Qwen3:4b exhibits significant reasoning limi–tations regarding complex control commands. While Qwen3:8b demonstrates high accuracy, its response timeisexcessive(70-98 seconds).Consequently, Llama 3.1:8b is identified as the optimal solution for this study, as it effectively balances processing speed with semantic reasoning capabilities, maintaining a stable average response time of under 10 seconds.

To mitigate invalid command generationand en–hance reliability, a structured prompt is designed to direct the LLM's reasoning flow. This study implements a functioncalling mechanism via the LangChain frame–workfor communication between the LLM and the drone[39]. The prompt integrates detailed functional descriptions and required parameters for each control function, establishing safety constraints on flight range, command frequency, and operational conditions to prevent actions that could jeopardize the hardware.

Specifically, theuser's input statement$X$ will pass through the loop's preprocessor support functions, denoted as$H_{aux}$to create a standardized input chain $X' = H_{aux}\left(X\right)$. Let's assume the output of the PhoWhisper speech recognition model is a sequence of$n$ words, represented as an ordered list $W = (w_1,w_2,\ldots,w_n)$ The repetitive word processing function $H_{aux}(W)$ performs an iteration to eliminate adjacent duplicate tokens (e.g., transforming the string "fly fly fly up" into "fly up"). This process is represented via the string concatenation operator as follows[40]:

$$W' = \oplus_{i=1}^{n} \delta\left(w_i, w_{i-1}\right) \tag{2}$$

where the conditional function δ is defined as:

$$\delta\left(w_i, w_{i-1}\right) = \begin{cases} w_i, & if \quad i=1 \quad or \quad w_i \neq w_{i-1} \\ \varepsilon & if \quad w_i = w_{i-1} \end{cases} \tag{3}$$

with ε is empty string. By applying this formula, the noisy data sequence $W$is filtered into a clean text sequence W' prior to LLM intent analysis. This process substantially reduces redundant tokens and prevents contextual misinterpretation by the model.

A physical action space Acomprises the following predefined drone control API functions:

$$A = \left\{takeoff(d), move(dir), rot(\theta), flip(dir), land\right\} \tag{4}$$

where:d and θ represent distance and rotation angle parameters, respectively.

Based on the sequence X' and the system prompt$P$, the LLM (parameterized by θ_{LLM}) performs intent ana– lysis. If the request involves physical actions, the model activates a function-calling mechanism rather than ge– nerating standard text. This process is modeled as a constrained decoding task, wherein the LLM generates a sequence of execution functions $C=\left[c_1, c_2, ..., c_k\right]$ (where $c_i \in A$) to maximize the joint probability distribution [41]:

$$C^* = \arg\max_{C \subset A} \prod_{i=1}^{k} P\left(c_i \mid X', P, c_{1:i-1}; \theta_{LLM}\right) \tag{5}$$

Forsafety reasons, unsupported or ambiguous com– mands are not executed. If the input command cannot be mapped to any predefined function in the action space A, the system rejects the command and switches to a safe fallback state. Specifically, the drone remains hovering when it is already airborne, or remains idle when it is on the ground. This prevents undefined LLM outputs from being converted into unsafe actions.

By restricting the output sample space to the set$A$, the function-calling mechanism ensures that every $\hat{c}_i \in C^*$ represents a feasible physical command, there– by eliminating word repetition and invalid command generationin command generation. Therefore, the pro– posed system follows a closed-set command execution strategy: only predefined API functions are executable, while unsupported commands trigger a hover/idle fallback response. The system facilitates the processing of compound commands, addressing the limitations of singlecommand execution,invalid command issues prevalent in prior research[42].

3. HARDWARE SETUP & EXPERIMENTAL RESULTS

Utilizing the hardware architecture established in Figure 5, the system was empirically deployed for performance evaluation. Experiments were conducted in a controlled environment with low ambient noise to isolate the models core capabilities from external variables. It should be noted that this study did not systematically evaluate the effects of Vietnamese regional accents, speaking styles, or tone variations. The experiments were conducted under controlled low-noise conditions to primarily validate the LLM-based command mapping and function-calling mechanism.

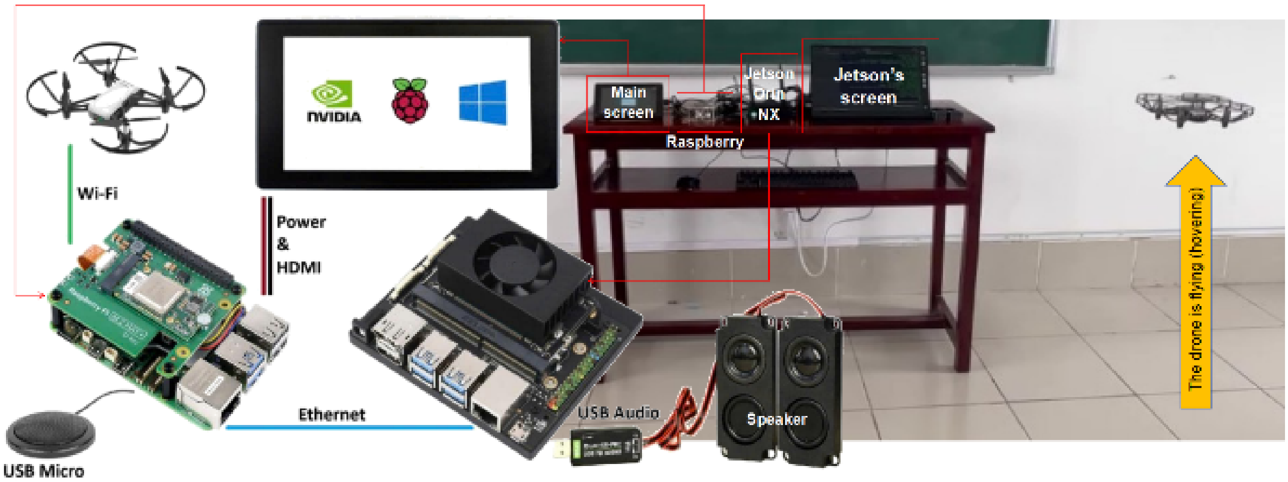

Figure 5. Hardware setup and system deployment in testing

The system is evaluated based on: (i) average res– ponse time per single command; (ii) command recog– nition and execution accuracy; and (iii) task execution precision. Testing tasks are categorized into three groups of increasing complexity single, compound, and complex (semantic reasoning) commands evaluated across various LLM architectures.A compound com– mand is defined as a sequence of multiple explicit drone actions stated in one query, whereas a complex com– mand requires additional semantic reasoning to infer the user's intent before mapping it to executable actions.

```
total duration:        2.081700315s
load duration:         124.788097ms
prompt eval count:     13 token(s)
prompt eval duration:  301.856976ms
prompt eval rate:      43.07 tokens/s
eval count:            13 token(s)
eval duration:         1.652796493s
eval rate:             7.87 tokens/s
>>> Send a message (/? for help)
```

Figure 6. LLM inference parameters on terminal interface

For the single-command group (Figure 6), the system demonstrated 100% accuracy with a mean processing and response latency of 2-3 seconds per command. Performance evaluation data for a single model query, extracted and displayed via the terminal interface (Figure 6), indicates a mean total duration of approximately 2.08 seconds, inclusive of model loading time. The evaluation rate on the hardware configuration (as specified in Figure 6) reached 7.87 tokens/s, a sufficient performance level for practical control appli- cations. This was further validated by the successful physical execution of the subsequent drone actions.

Table 1. Compound command scenario requiring complex reasoning: "Ascend 50 cm, hover for 5 seconds, perform a left flip, execute a 360-degree clockwise rotation, hover for 2 seconds, and land."

Model	Responsetime	Tokens output	Tokens/s
Qwen3:4b	N/A	N/A	N/A
Qwen3:8b	49.01 s	183	8.46
Llama3.1:8b	8.45 s	14	6.74

Table 2. Repetitive word detection and elimination scenario in compound command execution: “Fly fly fly up 50 cm, hover for 5 seconds, and land.”

Model	Response time	Tokens output	Tokens/s
Qwen 3:4b	76.72	594	11.91
Qwen 3:8b	21.25	147	7.78
Llama 3.1:8b	18.66	12	7.39

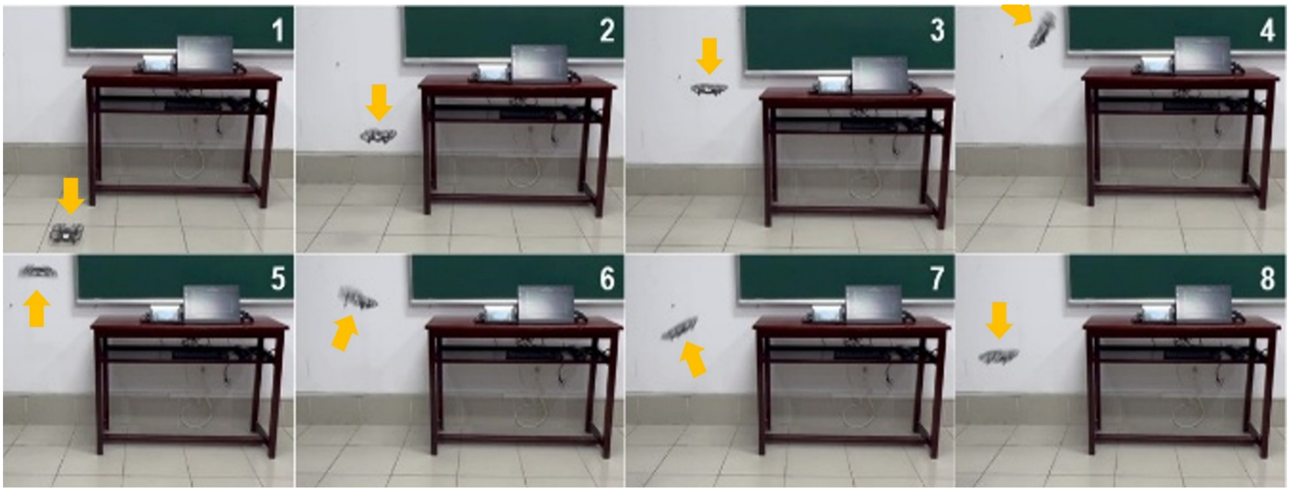

Figure 7. Physical drone execution in the scenario: “ascend 50 cm, perform a left flip, hover for 5 seconds, and land”

Experimental results summarized in Tables 1 and 2 reveal significant performance disparities among the models during UAV control command processing. Llama 3.1:8b demonstrated superior reasoning capabi–lities and format adherence. In both scenarios, this mo–del achieved accurate command extraction with a minimal output (12-14 tokens), indicating direct map–ping into the command space without generating redun–dant text (invalid command generation). Conversely, Qwen 3:4b failed to execute complex commands and proved highly susceptible to repetitive input noise (generating 594 tokens). Although Qwen 3:8b comp–leted the tasks, it produced excessive extraneous data (147-183 tokens). Precise output format control allowed Llama 3.1:8b to achieve the lowest total response latency (8.45 s in the reasoning scenario). While Qwen 3:8b exhibited a slightly higher token generation rate, the generation of redundant tokens resulted in prohibitive system latency (up to 49.01 s), rendering it unsuitable for near-real-time control. Under noisy input conditions featuring word repetition, Llama 3.1:8b maintained low latency and formatting accuracy, whereas the Qwen models exhibited anomalous processing delays.

These findings demonstrate that Llama 3.1:8b is the optimal model for the proposed system architecture. This model effectively integrates complex semantic reasoning, high robustness against word repetition noise, and strict adherence to output formats. Such characteristics eliminate redundant text generation, thereby minimizing end-to-end latency to meet UAV operational requirements.

Figure 7 illustrates drone behaviour in a compound command scenario: “ascend 50 cm, perform a left flip, hover for 5 seconds, and land,” which was processed and executed with precision. Under environmental scenarios with significant ambient noise, action success rates for single commands declined to 60-85%. For complex reasoning tasks, accuracy fell below 50%, and processing and inference latencies extended to several minutes.Although the proposed system achieved high command execution accuracy, latency remains a critical issue for practical UAV operation. The main sources of latency include speech recognition on the Raspberry Pi 5, LLM inference on the Jetson Orin NX, and the number of output tokens generated during function calling. The results in Tables 1 and 2 indicate that excessive token generation significantly increases response time. Therefore, latency can be improved by using lighter or more task-specific LLMs, reducing unnecessary output tokens through stricter function-calling constraints, and deploying the inference module on more powerful edge-computing hardware.

Overall, the system demonstrated stable performance within controlled low-noise environments. The end-to-end pipelinecomprising audio acquisition, speech recognition, LLM inference, and physical command executionoperated seamlessly, with nosystem interruptions observed under experimental conditions.For the compound and complex command categories requiring deep reasoning, the system achieved 100% accuracy, successfully performing semantic parsing and sequential action deployment within a single query.

4. CONCLUSION AND FUTURE WORKS

This study presents the design and implementation of a Vietnamese voice-controlled drone system based on the integration of LLM and distributed computing architecture. The system achieved 100% accuracyfor both single and sequential compound commands within a noise-controlled experimental environment.

The primary contribution of this study is the proposal of a structured prompt framework, enabling the LLM to perform precise and reliable natural language parsing, extraction, and function calling. This framework facilitates the accurate execution of complex, multi-step command sequences, overcoming the limi–

tations of preceding systems restricted to single-task processing per instruction.

Nevertheless, the system requires enhanced reasoning for metaphorical contexts, and its sensitivity to environmental noise remains a critical challenge. Consequently, future research will focus on improving context awareness and addressing metaphorical instructions that necessitate complex geometric trajectory calculations (e.g., square or planar circular patterns), alongside the integration of high-precision spatial positioning techniques. Moreover, future work will evaluate the robustness of the system using a more diverse Vietnamese speech dataset, including different regional accents, speaking styles, and speaker profiles.

REFERENCES

[1] M. Miloš and P. Mirosavljević, "Analysis of The Performance and Kinematics of The Movement of UAV," *FME Transactions,* vol. 51, no. 4, pp. 627-636, 2023. https://doi.org/10.5937/fme2304627M

[2] K. Kanistras, G. Martins, M. J. Rutherford, and K. P. Valavanis, "A survey of unmanned aerial vehicles (UAVs) for traffic monitoring," *2013 International Conference on Unmanned Aircraft Systems (ICUAS),* pp. 221-234, 2013. https://doi.org/10.1109/ICUAS.2013.6564694

[3] M. Erdelj, E. Natalizio, "UAV-assisted disaster management: Applications and open issues," presented at the 2016 International Conference on Computing, Networking and Communications (ICNC), 2016. https://doi.org/10.1109/iccnc.2016.7440563

[4] K. Dorling, J. Heinrichs, G. G. Messier, S. Magierowski, "Vehicle Routing Problems for Drone Delivery," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 47, no. 1, pp. 70-85, 2017. https://doi.org/10.1109/tsmc.2016.2582745

[5] X. Mei, F. Xu, Z. Zhang, Y. Li, "Unmanned Aerial Vehicle-Assisted Inspection for Transportation Infrastructure Operation and Maintenance: A Review," presented at the 2025 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2025. https://doi.org/10.1109/ieem63636.2025.11357603

[6] G. Pajares, "Overview and Current Status of Remote Sensing Applications Based on Unmanned Aerial Vehicles (UAVs)," Photogrammetric *Engineering & Remote Sensing,* vol. 81, no. 4, pp. 281-330, 2015. https://doi.org/10.14358/pers.81.4.281

[7] S. Zhang, X. Wang, H. Lin, Y. Dong, and Z. Qiang, "A review of the application of UAV multispectral remote sensing technology in precision agriculture," *Smart Agricultural Technology,* vol. 12, 2025. https://doi.org/10.1016/j.atech.2025.101406

[8] G. Ghazi and J. Voyer, "Use of a dji tello drone as an educational platform in the field of control engineering," *Espace ETS Publishcation,* 2024.

[9] M. Polishchuk, O. Rolik, "Improvement of technological equipment drone for water sampling: Design and modeling," *FME Transactions,* vol. 52, no. 2, pp. 237-245, 2024. https://doi.org/10.5937/fme2402237P

[10] N. M. Thamrin, N. H. M. Arshad, R. Adnan, R. Sam, N. A. Razak, M. F. Misnan, S. F. Mahmud, "Simultaneous localization and mapping based real-time inter-row tree tracking technique for unmanned aerial vehicle," presented at the 2012 IEEE International Conference on Control System, Computing and Engineering, 2012. https://doi.org/10.1109/iccsce.2012.6487164

[11] L. A. Fagundes-Junior, K. B. de Carvalho, R. S. Ferreira, A. S. Brandão, "Machine Learning for Unmanned Aerial Vehicles Navigation: An Overview," *SN Computer Science,* vol. 5, no. 2, 2024. https://doi.org/10.1007/s42979-023-02592-5

[12] D. Floreano, R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature,* vol. 521, no. 7553, pp. 460-6, May 28 2015. https://doi.org/10.1038/nature14542

[13] P. Garg, "Adaptive Multi-UAV Control Algorithm by Using Deep Learning," presented at the 2023 International Conference on Data Science and Network Security (ICDSNS), 2023. https://doi.org/10.1109/icdsns58469.2023.10245971

[14] J. R. Cauchard, J. L. E, K. Y. Zhai, J. A. Landay, "Drone & me: an exploration into natural human-drone interaction," presented at the Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2015. https://doi.org/10.1145/2750858.2805823

[15] A. D. Abhay, M. A. P. Dr, "A Review Study on Applications of Natural Language Processing," *International Journal of Scientific* Research *in Science, Engineering and Technology,* pp. 122-126, 2023. https://doi.org/10.32628/ijsrset2310214

[16] Y. Feng, H. Snoussi, J. Teng, A. Cherouat, and T. Wang, "Large language model-based multi-task UAVs - towards distilled real-time interactive control," *IET Conference Proceedings,* vol. 2023, no. 39, pp. 114-118, 2024. https://doi.org/10.1049/icp.2024.0472

[17] H. Li, M. Xiao, K. Wang, D. I. Kim, M. Debbah, "Large Language Model Based Multi-Objective Optimization for Integrated Sensing and Communications in UAV Networks," *IEEE Wireless Communications Letters,* vol. 14, no. 4, pp. 979-983, 2025. https://doi.org/10.1109/lwc.2025.3529082

[18] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys,* vol. 55, no. 12, pp. 1-38, 2023. https://doi.org/10.1145/3571730

[19] T. S. Gurunathan, M. S. Raza, A. K. Janakiraman, M. A. Khan, B. Pal, A. *Gangopadhyay*, "Edge LLMs for Real-Time Contextual Understanding with Ground Robots," *Proceedings of the AAAI Symposium Series,* vol. 5, no. 1, pp. 159-166, 2025. https://doi.org/10.1609/aaaiss.v5i1.35583

[20] Y. Zhai, Y.-T. Chian, "Edge-Deployed Context-Aware LLM Framework for Low-*Latency* Bi-Directional

Gaze-Speech HRI," *Preprints,* p. 2026020308, 2026. https://doi.org/10.20944/preprints202602.0308. v1

[21] T. Kollar, S. Tellex, D. Roy, N. Roy, "Toward *understanding* natural language directions," presented at the Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10, 2010. https://doi.org/10.1145/1734454. 1734553

[22] A. Lytovchenko, "Voice-Based Control of Unmanned Aerial Vehicles using Natural Language Processing and Machine Learning Methods," presented at the 2025 IEEE 8th International Conference on Methods and Systems of *Navigation* and Motion Control (MSNMC), 2025. https://doi.org/10.1109/ msnmc61130.2025.11399713

[23] C. K. Singh, D. Kumar, V. Sanap, R. Sinha, "LLM-RSPF: Large Language Model-Based Robotic *System* Planning Framework for Domain Specific Use-cases," presented at the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025. https://doi.org/10.1109/wacv 61041.2025.00707

[24] Y. Alkasim, A. Altahhan, "Enhancing Voice-Controlled Drone Navigation: A Hybrid Approach Using ASR and NLP for UAV Command Interpretation," in *Artificial Intelligence XLII*, Cham, M. Bramer and F. Stahl, Eds., 2026 2026: Springer Nature Switzerland, pp. 253-269. https://doi.org/ 10.1007/978-3-032-11442-6_18

[25] L. Yuan, C. Deng, D.-J. Han, I. Hwang, S. Brunswicker, and C. G. Brinton, "Next-Generation LLM for UAV: From Natural Language to Autonomous Flight," *ArXiv,* vol. abs/2510.21739, 2025. https://doi.org/10.48550/arXiv.2510.21739

[26] P. García, L. Arribas, P. Pueyo, L. Riazuelo, A. C. Murillo, "Exploring dji tello as an affordable drone solution for research and education," in *2024 7th Iberian Robotics Conference (ROBOT)*, 2024: IEEE, pp. 1-6. https://doi.org/10.1109/ROBOT 61475.2024.10796954

[27] F. Candan, O. F. Dik, T. Kumbasar, M. Mahfouf, L. Mihaylova, "Real-time interval type-2 fuzzy control of an unmanned aerial vehicle with flexible cable-connected payload," *Algorithms,* vol. 16, no. 6, p. 273, 2023. https://doi.org/10.3390/a16060273

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, "Robust speech recognition via large-scale weak supervision," presented at the Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA, 2023.

[29] T.-T. Le, L. T. Nguyen, D. Q. Nguyen, "Phowhisper: Automatic speech recognition for vietnamese," *arXiv preprint arXiv:2406.02555,* 2024. https://doi.org/10.48550/arXiv.2406.02555

[30] Y. Cao, "Performance Evaluation of Whisper-Series Speech Transcription Models on Raspberry Pi," in *Proceedings of the Tenth ACM/IEEE Symposium on Edge Computing*, 2025, pp. 1-7. https://doi.org/ 10.1145/3769102.3774244

[31] S. Sarkar, M. F. Babar, M. M. Hassan, M. Hasan, S. K. Karmaker Santu, "Processing natural language on embedded devices: How well do transformer models perform?," in *Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering*, 2024, pp. 211-222. https:// doi.org/10.1145/3629526.3645054

[32] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38-45. https://doi.org/10.18653/v1/2020.emnlpdemo s.6

[33] W. Chan, N. Jaitly, Q. Le, O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech *recognition*," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2016: IEEE, pp. 4960-4964. https://doi.org/10.1109/ICASSP.2016. 7472621

[34] J. Kim, J. Kong, J. Son, "Conditional variational autoencoder with adversarial *learning* for end-to-end text-to-speech," in *International conference on machine learning*, 2021: PMLR, pp. 5530-5540.

[35] T. Keith, "Work in Progress: Text-to-speech on Edge Devices for te Reo Māori and ʻŌlelo Hawai□i," in *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, 2024, pp. 421-426.

[36] M. Ahn *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691,* 2022. https://doi.org/10.48550 /arXiv.2204.01691

[37] A. Grattafiori *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783,* 2024. https://doi.org/10.48550/arXiv.2407.21783

[38] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, K. Huang, "Pushing large language models to the 6g edge: Vision, challenges, and opportunities," *IEEE Communications Magazine,* vol. 63, no. 9, pp. 52-59, 2025. https://doi.org/10.1109/MCOM.001. 2400764

[39] S. G. Patil, T. Zhang, X. Wang, J. E. Gonzalez, "Gorilla: Large language model connected with massive apis," *Advances in Neural Information Processing Systems,* vol. 37, pp. 126544-126565, 2024. https://doi.org/10.52202/079017-4020

[40] P. J. Lou, M. Johnson, "End-to-end speech recognition and disfluency removal," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2051-2061. https://doi.org/10. 18653/v1/2020.findings-emnlp.186

[41] T. Schick *et al.*, "Toolformer: Language models can teach *themselves* to use tools," *Advances in neural information processing systems,* vol. 36, pp. 68539-68551, 2023.

[42] M. Al Khawaldah, O. Daoud, S. S. Alja'afreh, A. M. Matarneh, "GPT Voice-Driven Natural Language Control for ArduPilot-Guided *UAVs*," in *2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD)*, 2025: IEEE, pp. 296-300. https://doi.org/10.1109/SSD64182.2025.10990012

NOMENCLATURE & ABBREVIATIONS

UAV	Unmanned aerial vehicle
LLM	Large language model
NLP	Natural language processing
API	Application programming interface
ASR	Automatic speech recognition
I/O	Input/output
GPU	Graphics processing unit
CPU	Central processing unit
GUI	Graphical user interface
LAN	Local area network
IP	Internet protocol
USB	Universal serial bus

КОНТРОЛА ДРОНА ЗАСНОВАНА НА LLMANDNLP-У: АРХИТЕКТУРА ДИСТРИБУИРАНОГ ИВИЧНОГ РАЧУНАРСТВА ЗА ВИЈЕТНАМЦЕ

Д.Т. Нгујен, Ф.Л. Тран, Ч.Н. Нгујен, Х.Ф. Тран, Д.Ф. Тран

Ова студија предлаже вијетнамски систем гласовне контроле заснован на ивици за беспилотне летелице. Систем интегрише велики језички модел, модуле за обраду природног језика и стварно активирање дроном како би се побољшала поузданост управ–љања дроном заснованом на природном језику. Главни допринос је уграђени мехатронички конт–ролни оквир експериментално потврђен на комер–цијалној платформи за дрон. Овај оквир комбинује вишеслојну структуру промпта са механизмом за претходну обраду да би се смањила реч која се понавља, неважеће генерисање команди и редун–дантни текстуални излаз. Предложени систем ко–ристи архитектуру дистрибуиране ивице. NVIDIA Jetson Orin NKS изводи закључивање Llama 3.1:8б, док Распберрихендс користи препознавање говора засновано на ПхоВхисперу и директну DJI Tello контролу. У условима контролисаним буком, систем је постигао 100% тачност за појединачне, сложене и инференцијалне командне категорије. Просечно вр–еме одговора било је 2–3 с за појединачне команде и 8,45 с за репрезентативну сложену сложену команду.