Nava : Vers une Architecture Souveraine et Miniaturisée pour Agents Conversationnels

Nima FATHOLLAHZADEH - nima@vigilantia.fr

Février 2025

1 Résumé et Contexte

Nava est une plateforme dintelligence artificielle conversationnelle conçue pour accueillir vos clients par téléphone, à toute heure, et enregistrer automatiquement leurs réservations. Toutes les informations sont synchronisées en temps réel sur notre application web et mobile, garantissant une mise à jour instantanée des plannings. En combinant une reconnaissance vocale (STT) basée sur NeMo ASR, un modèle de langage (LLM) dérivé de Llama finement adapté et une synthèse vocale (TTS) via NeMo TTS, Nava forme une solution robuste et ultra-optimisée, idéale pour fonctionner même sur des dispositifs embarqués à ressources limitées.

Notre approche repose sur un fine-tuning intensif de chaque module pour ladapter à votre secteur et vos spécificités, ainsi que sur une miniaturisation poussée des modèles grâce à des techniques de quantification et de pruning. Cette double stratégie nous permet datteindre un haut niveau de performance, tout en assurant la souveraineté totale de vos données et en offrant un service fiable, réactif et économique.

2 Travaux de Référence

Les recherches récentes sur l'IA conversationnelle identifient des leviers clés pour son déploiement dans l'HORECA. L'ingénierie des prompts et les boucles de rétroaction sur la qualité des interactions améliorent la précision des réponses des modèles de langage. Les interfaces vocales, perçues comme plus efficaces que les chatbots textuels, réduisent la charge cognitive lorsque l'objectif est explicite. Les progrès en reconnaissance vocale (STT) permettent des interactions fluides, approchant la précision humaine. Des indicateurs actionnables (clarté de l'intention, contexte) optimisent les dialogues spécialisés. Les agents multilingues basés sur LLM et l'intégration de données géospatiales améliorent l'adaptabilité opérationnelle tout en limitant la dépendance au cloud. Ces avancées soulignent l'importance des interfaces multimodales, des architectures edge et des mécanismes d'auto-optimisation pour les solutions d'intelligence artificielle nouvelle génération dans l'hôtellerie.

3 Introduction

Lévolution rapide de lintelligence artificielle a ouvert des perspectives inédites pour automatiser la communication et les interactions clients. Toutefois, la dépendance aux solutions étrangères et la lourdeur des modèles actuels posent des défis en termes de souveraineté des données, de respect de la vie privée et doptimisation sur du matériel embarqué. Nava sinscrit dans cette dynamique en proposant une architecture entièrement locale et fine-tunée, permettant dexploiter les technologies de pointe en STT, LLM et TTS dans un format miniaturisé, adapté aux contraintes matérielles des dispositifs tels que le Jetson Orin.

4 Défis et Enjeux

4.1 Souveraineté des données

Garantir que lensemble du traitement (de la reconnaissance vocale à la synthèse) se fasse en local, sans dépendre dinfrastructures externes propriétaires et centralisées.

4.2 Optimisation sur dispositifs embarqués

Réduire la taille des modèles par quantification et pruning pour assurer une exécution en temps réel sur des plateformes à ressources limitées, telles que le Jetson Orin de Nvidia.

4.3 Adaptabilité linguistique

Adapter finement les modèles aux spécificités du français et aux accents locaux pour une reconnaissance et une génération de texte optimales.

4.4 Intégration et interopérabilité

Mettre en place une solution modulaire capable de sintégrer dans divers écosystèmes métier, notamment pour la gestion des rendez-vous et la planification, tout en restant évolutive.

5 Architecture Technique

Notre système intègre trois composantes majeures :

5.1 Reconnaissance Vocale (STT)

Cette composante convertit la parole en texte grâce à des modèles optimisés par quantification et pruning pour garantir une latence minimale sur des dispositifs embarqués.

5.2 Modèle de Langage (LLM)

En utilisant un LLM libre comme socle, nous appliquons un fine-tuning intensif pour adapter le modèle aux spécificités du français et des secteurs ciblés, tout en le miniaturisant (quantification) afin de réduire son empreinte.

5.3 Synthèse Vocale (TTS)

Sappuyant sur les algorithmes FastPitch + HiFi-GAN, cette composante transforme le texte généré par le LLM en une parole naturelle et expressive, optimisée pour une exécution sur matériel limité sans perte de qualité sonore.

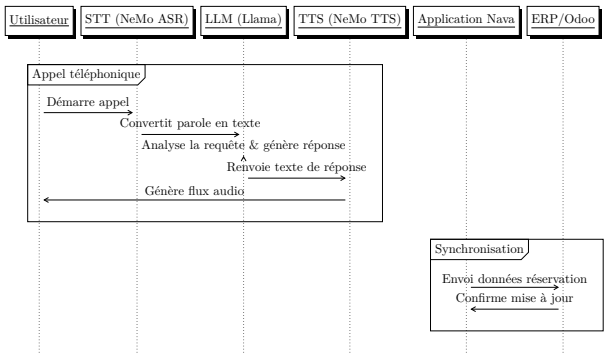

FIGURE 1 – Diagramme de séquence Interaction utilisateur, STT, LLM, TTS et synchronisation avec ERP

6 Agent IA fonctionnels

Les agents IA de Nava sont conçus pour mener des conversations structurées, analyser les requêtes et générer des réponses contextuelles adaptées. Chaque agent opère sur la base de connaissances spécifiques au secteur, assurant ainsi une compréhension fine des interactions et une exécution précise des tâches.

7 Miniaturisation

Pour garantir une exécution rapide et une empreinte mémoire réduite, nous appliquons trois techniques clés. **Quantification** en 4-bit ou INT8 limite la consommation de RAM sans nuire de façon significative aux performances. **Pruning** élimine les couches redondantes pour réduire la complexité du modèle tout en préservant sa précision. Enfin, l**optimisation TensorRT** assure une exécution hautement performante sur des plateformes comme le Jetson Orin, notamment pour les modules ASR et TTS.

8 Fine-tuning

Nous réalisons un fine-tuning approfondi afin dadapter les modèles aux besoins sectoriels (ex. restauration, services) et aux spécificités linguistiques, y compris les accents locaux. Des jeux de données dédiés sont développés pour chaque domaine et région, et des techniques de fine-tuning ciblées garantissent une compréhension et une génération parfaitement alignées sur le contexte dutilisation.

9 Intégration dans lÉcosystème Partenaires et ERP

Nava est conçue pour sintégrer facilement dans des environnements collaboratifs et interconnectés. Grâce à une architecture modulaire, elle communique aisément avec des systèmes tiers, dont lécosystème Odoo pour la gestion fluide des rendez-vous et du planning. Cette intégration assure une synchronisation efficace avec les systèmes ERP et SMS, offrant ainsi une solution complète et parfaitement interopérable.

10 Appel à Collaboration

Nava incarne lavenir de lintelligence artificielle conversationnelle souveraine. Nous recherchons activement des partenaires stratégiques, des investisseurs et des talents passionnés par linnovation et la souveraineté des données. Rejoignez-nous pour transformer lIA conversationnelle et offrir aux entreprises une solution véritablement décentralisée et optimisée.

11 Références

— Meysam Shamsi. Speech Emotion Classification from Affective Dimensions : Limitation and Advantage. 11th International Conference on Affective Computing and Intelligent Interaction (ACII), MIT Media Lab, Jul 2023, Cambridge Massachusetts, USA, France. pp.1-4, ff10.1109/ACIIW59127.2023.10388084ff. ffhal-04239244f `https://hal.science/hal-04239244v1`

— Towards a Conversational LLM-Based Voice Assistant for Transportation Applications S Jafarnejad, A Berthe-Pardo, R Frank - 2024 IEEE Vehicular Networking Conference (VNC), 2024 `https://orbilu.uni.lu/bitstream/10993/61643/1/2024126343.pdf`

— Eva : An LLM-based Multilingual Voice-agent Network for Restaurant Operations Zhiwei (Tony) Qin1, Jianming Zhou1, 1Eva AI `https://s3.us-west-2.amazonaws.com/tpass.media/eva/ic_eva_white_paper.pdf`