Une architecture multi-agents pour la génération automatique de tickets en environnement industriel : focus sur l'agent de classification

Ying ZHANG [1], Sébastien BONNET [2], Matthieu PETIT GUILLAUME [1], Muriel HUG [1], Aurélien KRAUTH [1], Rémi UHARTEGARAY [2]

1. Leviatan, 2. AntemetA

APIA 2025 - Dijon
01/07/2025

LEVIATAN ANTEMETA

LEVIATAN ANTEMETA

Problématique industrielle - Un besoin d'automatisation fiable

- **Volume croissant d'e-mails**
  - Un volume important de messages est traité quotidiennement

LEVIATAN ANTEMETA

Problématique industrielle - Un besoin d'automatisation fiable

- **Volume croissant d'e-mails**
  - Un volume important de messages est traité quotidiennement
- **Difficulté de classification** :
  - Un nombre élevé de catégories nécessitant une catégorisation fine
  - Nécessite une connaissance approfondie des catégories par les agents humains
- **Impact direct sur la qualité du service client**
  - Mauvaise catégorisation $\Rightarrow$ mauvaise affectation
  - Charge de travail accrue pour les opérateurs

LEVIATAN ANTEMETA

Problématique industrielle - Un besoin d'automatisation fiable

- **Volume croissant d'e-mails**
  - Un volume important de messages est traité quotidiennement
- **Difficulté de classification** :
  - Un nombre élevé de catégories nécessitant une catégorisation fine
  - Nécessite une connaissance approfondie des catégories par les agents humains
- **Impact direct sur la qualité du service client**
  - Mauvaise catégorisation $\Rightarrow$ mauvaise affectation
  - Charge de travail accrue pour les opérateurs
- **Besoins métiers**
  - Automatiser le traitement complet des demandes clients
  - Garantir un système cohérent et fiable
  - Intégrer automatiquement les tickets dans les outils internes

LEVIATAN ANTEMETA

Contraintes spécifiques - Enjeux industriels

- **Confidentialité des données**
  - Les e-mails contiennent des informations sensibles
  - Interdiction d'utiliser des services publics ou cloud externes

LEVIATAN ANTEMETA

Contraintes spécifiques - Enjeux industriels

- **Confidentialité des données**
  - Les e-mails contiennent des informations sensibles
  - Interdiction d'utiliser des services publics ou cloud externes
- **Contraintes matérielles**
  - Les modèles de langage de grande taille nécessitent des GPU puissants et onéreux
  - Souvent incompatible avec l'infrastructure informatique typique d'une PME

LEVIATAN ANTEMETA

Contraintes spécifiques - Enjeux industriels

- **Confidentialité des données**
  - Les e-mails contiennent des informations sensibles
  - Interdiction d'utiliser des services publics ou cloud externes
- **Contraintes matérielles**
  - Les modèles de langage de grande taille nécessitent des GPU puissants et onéreux
  - Souvent incompatible avec l'infrastructure informatique typique d'une PME
- **Exigence de fiabilité du système**
  - Le système doit produire des résultats cohérents et exploitables
  - Les erreurs peuvent impacter la prise en charge des demandes critiques

LEVIATAN ANTEMETA

Principe multi-agents - Une architecture modulaire

- **Un système composé de plusieurs agents spécialisés**
  - **Agent de classification** : identifie la catégorie de l'e-mail
  - **Agents d'extraction** : extraient les champs nécessaires selon la catégorie
  - **Orchestrateur** : coordonne les agents et pilote le flux

LEVIATAN ANTEMETA

Principe multi-agents - Une architecture modulaire

- **Un système composé de plusieurs agents spécialisés**
  - **Agent de classification** : identifie la catégorie de l'e-mail
  - **Agents d'extraction** : extraient les champs nécessaires selon la catégorie
  - **Orchestrateur** : coordonne les agents et pilote le flux
- **Pourquoi une architecture multi-agents ?**
  - Simplifie chaque tâche $\Rightarrow$ prompts plus courts et plus ciblés
  - Une meilleure lisibilité et précision pour chaque agent
  - Les agents d'extraction peuvent être ajoutés ou modifiés indépendamment, sans affecter les autres

LEVIATAN ANTEMETA

Principe multi-agents - Une architecture modulaire

- **Un système composé de plusieurs agents spécialisés**
  - **Agent de classification** : identifie la catégorie de l'e-mail
  - **Agents d'extraction** : extraient les champs nécessaires selon la catégorie
  - **Orchestrateur** : coordonne les agents et pilote le flux
- **Pourquoi une architecture multi-agents ?**
  - Simplifie chaque tâche $\Rightarrow$ prompts plus courts et plus ciblés
  - Une meilleure lisibilité et précision pour chaque agent
  - Les agents d'extraction peuvent être ajoutés ou modifiés indépendamment, sans affecter les autres
- **Approche progressive**
  - Le projet se concentre d'abord sur l'agent de classification
  - Les agents d'extraction seront développés par la suite

LEVIATAN ANTEMETA

LEVIATAN ANTEMETA

Agent de classification - Objectifs et choix du modèle

- **Objectifs**
  - Identifier la catégorie adéquate pour chaque e-mail
  - S'intégrer facilement à l'orchestrateur multi-agents

Agent de classification - Objectifs et choix du modèle

- **Objectifs**
  - Identifier la catégorie adéquate pour chaque e-mail
  - S'intégrer facilement à l'orchestrateur multi-agents
- **Contraintes industrielles**
  - Infrastructure limitée : 1 GPU NVIDIA A10 (24 Go VRAM)
  - Confidentialité : déploiement on-premise, aucun appel externe
  - Robustesse : une exactitude élevée est indispensable en production

LEVIATAN ANTEMETA

Agent de classification - Objectifs et choix du modèle

- **Objectifs**
  - Identifier la catégorie adéquate pour chaque e-mail
  - S'intégrer facilement à l'orchestrateur multi-agents
- **Contraintes industrielles**
  - Infrastructure limitée : 1 GPU NVIDIA A10 (24 Go VRAM)
  - Confidentialité : déploiement on-premise, aucun appel externe
  - Robustesse : une exactitude élevée est indispensable en production
- **Choix du modèle**
  - LLM open-source d'environ 13 milliards de paramètres
  - Quantification 4 bits pour réduire l'empreinte mémoire [4]

LEVIATAN ANTEMETA

Constitution du jeu de données - Nettoyage et sélection

- **Source et volume brut**
  - Données issues de la base réelle de gestion des tickets
  - 175 813 lignes issus de la base de support
  - Problèmes identifiés :
    - Plus de 200 catégories contenaient moins de 5 lignes
    - Un même ticket peut générer plusieurs lignes

Constitution du jeu de données - Nettoyage et sélection

- **Source et volume brut**
  - Données issues de la base réelle de gestion des tickets
  - 175 813 lignes issus de la base de support
  - Problèmes identifiés :
    - Plus de 200 catégories contenaient moins de 5 lignes
    - Un même ticket peut générer plusieurs lignes
- **Nettoyage et prétraitement**
  - Filtrage des colonnes nécessaires $\Rightarrow$ 154 414 lignes
  - Regroupement par *Numéro* $\Rightarrow$ 29 376 tickets uniques
  - Extraction du premier e-mail de chaque discussion
  - Sélection des catégories
    - 365 catégories initiales $\Rightarrow$ 9 retenues
    - Critères : impact métier et fréquence $\geq 2\,‰$

LEVIATAN ANTEMETA

Constitution du jeu de données - Nettoyage et sélection

- **Source et volume brut**
  - Données issues de la base réelle de gestion des tickets
  - 175 813 lignes issus de la base de support
  - Problèmes identifiés :
    - Plus de 200 catégories contenaient moins de 5 lignes
    - Un même ticket peut générer plusieurs lignes
- **Nettoyage et prétraitement**
  - Filtrage des colonnes nécessaires $\Rightarrow$ 154 414 lignes
  - Regroupement par *Numéro* $\Rightarrow$ 29 376 tickets uniques
  - Extraction du premier e-mail de chaque discussion
  - Sélection des catégories
    - 365 catégories initiales $\Rightarrow$ 9 retenues
    - Critères : impact métier et fréquence $\geq 2\,‰$
- **Jeu final pour l'apprentissage**
  - 4 500 tickets répartis en 9 catégories
  - 80 % train / 20 % test

LEVIATAN ANTEMETA

Constitution du jeu de données - Répartition finale des tickets (Train / Test)

Catégorie	**Train**	**Test**
Incidents/Supervision	2 397	565
Demande de service/Backup #BCS/Restauration qualifiée	361	83
Incidents/Backup #BCS/Sauvegarde	346	95
Demande de service/Backup #BCS/Autre	200	52
Demande de service/Backup #BCS/Stratégie de sauvegarde/-Création	78	21
Demande de service/Backup #BCS/Stratégie de sauvegarde/-Suppression	69	25
Demande de service/Backup #BCS/Demande de renseignement	58	17
Demande de service/Backup #BCS/Stratégie de sauvegarde/-Modification	50	14
Demande de service/Cyber Sécurité #CS2/Bastion/Création-Modification d'entrées	49	20
Total	**3 608**	**892**

$\Longrightarrow$ 80 % train / 20 % test, soit 4 500 tickets au total

LEVIATAN ANTEMETA

LEVIATAN ANTEMETA

Benchmark initial - Plateformes et méthodologie d'évaluation

- **Deux plateformes testées**
  - **Ollama** [10] : prompt JSON strict
    - Généré automatiquement via `pydantic.BaseModel` [1]
    - Schéma obtenu par `model_json_schema()`
  - **Unsloth** [3] : prompt en langage naturel

Benchmark initial - Plateformes et méthodologie d'évaluation

- **Deux plateformes testées**
  - **Ollama** [10] : prompt JSON strict
    - Généré automatiquement via `pydantic.BaseModel` [1]
    - Schéma obtenu par `model_json_schema()`
  - **Unsloth** [3] : prompt en langage naturel
- **Méthodologie d'évaluation**
  - **Matched vs Unmatched**
    - **Matched** : le LLM génère une catégorie claire
    - **Unmatched** : génération trop ambiguë, catégorie non identifiable
  - **Exactitude (accuracy)**

LEVIATAN ANTEMETA

Benchmark initial - Ingénierie des prompts

- Une explication de la tâche
- Une liste explicite des catégories possibles
- Des règles strictes de classification
- Un exemple annoté
- Une définition claire du format attendu (JSON ou texte libre)

```
# Define prompt for ticket classification assistant
prompt = """
Tu es un assistant spécialisé dans la classification de tickets à partir de le
```
Demande de service/Backup #BCS/Autre - Autre
Demande de service/Backup #BCS/Demande de renseignement - Demande de renseigne
Demande de service/Backup #BCS/Restauration qualifiée - Restauration qualifiée
Demande de service/Backup #BCS/Stratégie de sauvegarde/Création - Création
Demande de service/Backup #BCS/Stratégie de sauvegarde/Modification - Modifica
Demande de service/Backup #BCS/Stratégie de sauvegarde/Suppression - Suppressi
Demande de service/Cyber Sécurité #CS2/Bastion/Création-Modification d'entrées
Incidents/Backup #BCS/Sauvegarde - Sauvegarde
Incidents/Supervision - Supervision
```
### Règles :
1. Réponds uniquement par la catégorie exacte, sans texte supplémentaire.
2. La catégorie associée doit être unique.
3. Si aucune catégorie ne correspond, la valeur de "categorisation" doit être

### Exemple :
La description:
Bonjour¤ Merci de relancer les sauvegardes FULL¤ si ils ne sont pas repassées

### Sortie :
{{
  "categorisation": "Incidents/Backup #BCS/Sauvegarde - Sauvegarde"
}}

Analyse uniquement la description suivante :
{content}
"""
```

Exemple de prompt JSON

Benchmark initial - Plateformes et résultats

Modèle	Plateforme-Prompt	Mat./Unmat.	Match Rate	Exactitude
Llama 3.2 11B Vision [7]	Ollama-JSON	891/1	**0,9988**	0,8049
Qwen 2.5 14B [11]	Ollama-JSON	850/42	0,9529	**0,8520**
Mistral-Nemo 12B [9]	Ollama-JSON	659/233	0,7645	0,6446
Pixtral 12B [8]	Unsloth-Naturel	465/427	0,5213	0,4159
Qwen 2.5 14B	Unsloth-Naturel	523/369	0,5863	0,4776
Llama 3.2 11B Vision	Unsloth-Naturel	746/146	0,8363	0,4540

$\Longrightarrow$ *Les modèles testés sur **Ollama** obtiennent des performances nettement meilleures que sur Unsloth.*

LEVIATAN ANTEMETA

Fine-tuning - Contexte et méthode

- **Pourquoi fine-tuner ?**
  - Adapter le modèle aux catégories métier spécifiques
  - Améliorer la stabilité et la précision

Fine-tuning - Contexte et méthode

- **Pourquoi fine-tuner ?**
  - Adapter le modèle aux catégories métier spécifiques
  - Améliorer la stabilité et la précision
- **Modèles retenus** :
  - **Qwen 2.5 14B** : modèle purement linguistique
  - **LLaMA 3.2 11B Vision** : multimodal, mais image tower figée

Fine-tuning - Contexte et méthode

- **Pourquoi fine-tuner ?**
  - Adapter le modèle aux catégories métier spécifiques
  - Améliorer la stabilité et la précision
- **Modèles retenus** :
  - **Qwen 2.5 14B** : modèle purement linguistique
  - **LLaMA 3.2 11B Vision** : multimodal, mais image tower figée
- **Protocole d'entraînement** :
  - Plateforme : **Unsloth**
  - Approche utilisée : **PEFT avec LoRA** (Low-Rank Adaptation) [4, 6], avec quantification 4 bits
  - Deux variantes par modèle : **30 étapes** et **1 époque** ($\approx$ 440 étapes)
  - Annotation : Utilisation d'un **pseudo-JSON** basé sur des consignes en langage naturel
  - Jeu de données : Train set

LEVIATAN | ANTEMETA

Résultats des modèles ajustés

Modèle	Plateforme-Prompt	Mat./Unmat.	Match Rate	Exactitude
Qwen - 1 ep.	Ollama-JSON *	891/1	0,9989	0,8285
Qwen - 30 steps.	Ollama-JSON *	873/19	0,9787	0,8509
Qwen - 1 ep.	**Unsloth-Naturel (pseudo-JSON)**	**892/0**	**1,0**	**0,8812**
Qwen - 30 steps.	Unsloth-Naturel (pseudo-JSON)	886/6	0,9933	0,8565
Llama - 1 ep.	Unsloth-Naturel (pseudo-JSON)	889/3	0,9966	0,8610
Llama - 30 steps.	Unsloth-Naturel (pseudo-JSON)	870/22	0,9753	0,8105

* Conversion en format GGUF pour exécution sur Ollama.

LEVIATAN ANTEMETA

Comparaison des performances - Analyses

Observations clés

- → Qwen & LLaMA : amélioration des performances après fine-tuning
- → LLaMA : 0,4540-Unsloth/0,8049-Ollama ⇒ **0,8610**
- → Qwen : 0,4776-Unsloth/0,8520-Ollama ⇒ **0,8812**
- → **Top 0,8812** (Qwen, 1 époque, Unsloth, Pseudo-JSON)

LEVIATAN ANTEMETA

Comparaison des performances - Analyses

Observations clés

- → Qwen & LLaMA : amélioration des performances après fine-tuning
- → LLaMA : 0,4540-Unsloth/0,8049-Ollama ⇒ **0,8610**
- → Qwen : 0,4776-Unsloth/0,8520-Ollama ⇒ **0,8812**
- → **Top 0,8812** (Qwen, 1 époque, Unsloth, Pseudo-JSON)

Écart de performance observé pour Qwen sous Ollama - une hypothèse

- → Qwen : **0,8520** ⇒ **0,8285**
- → Quantification en NF4 via QLoRA durant le fine-tuning
- → Conversion en GGUF pour Ollama : reconstruction en FP16, puis re-quantification en Q4_K [2, 5]

LEVIATAN ANTEMETA

LEVIATAN ANTEMETA

Perspectives

À court terme

- **Optimiser l'agent de classification**
  - Ré-équilibrer et élargir le jeu de données
  - Re-fine-tuning du modèle *Qwen 2.5*
- **Développer les agents d'extraction**
  - PoC « Restauration qualifiée » : 5 champs clés
  - *Qwen 14B* et *Llama 3.2* (modèles de base, sans fine-tuning) : meilleurs scores
  - Finaliser l'annotation, puis fine-tuning ciblé

LEVIATAN ANTEMETA

Perspectives

À court terme

- **Optimiser l'agent de classification**
  - Ré-équilibrer et élargir le jeu de données
  - Re-fine-tuning du modèle *Qwen 2.5*
- **Développer les agents d'extraction**
  - PoC « Restauration qualifiée » : 5 champs clés
  - *Qwen 14B* et *Llama 3.2* (modèles de base, sans fine-tuning) : meilleurs scores
  - Finaliser l'annotation, puis fine-tuning ciblé

À long terme

- Unifier classification & extraction dans une architecture unique pour la génération et le suivi automatisés des tickets.

LEVIATAN ANTEMETA

LEVIATAN ANTEMETA

Références I

[1] Samuel Colvin et al. *Pydantic*. Version 1.10.20. 2025. DOI : 10.5281/zenodo.14615433. URL : https://github.com/pydantic/pydantic.

[2] Hugging Face Community. *Finetuned LLM Model Conversion to GGUF - Performance Drop*. https://discuss.huggingface.co/t/finetuned-llm-model-conversion-to-gguf-performance-drop/93179. Accessed : 2025-06-03. 2024.

[3] Michael Han Daniel Han et Unsloth team. *Unsloth*. 2023. URL : http://github.com/unslothai/unsloth.

[4] Tim Dettmers et al. *QLoRA : Efficient Finetuning of Quantized LLMs*. 2023. arXiv : 2305.14314 [cs.LG]. URL : https://arxiv.org/abs/2305.14314.

[5] Yixiao Li et al. "LoftQ : LoRA-Fine-Tuning-Aware Quantization for Large Language Models". In : *arXiv preprint arXiv :2310.08659* (2023). URL : https://arxiv.org/abs/2310.08659.

[6] Sourab Mangrulkar et al. *PEFT : State-of-the-art Parameter-Efficient Fine-Tuning methods*. 2022. URL : https://github.com/huggingface/peft.

[7] Meta. *Llama 3.2 : Revolutionizing Edge AI and Vision with Open, Customizable Models*. Accessed : 2025-02-21. Sept. 2024. URL : https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

[8] Mistral. *Announcing Pixtral 12B*. Accessed : 2025-02-21. Sept. 2024. URL : https://mistral.ai/news/pixtral-12b.

[9] Mistral. *Mistral NeMo*. Accessed : 2025-02-21. Juill. 2024. URL : https://mistral.ai/news/mistral-nemo.

[10] Ollama. *Ollama*. Accessed : 2025-02-21. URL : https://ollama.com/.

[11] An Yang et al. *Qwen2.5 Technical Report*. 2024.

LEVIATAN ANTEMETA

Merci !

Des questions ?

LEVIATAN ANTEMETA