AI Agents :Final Project Instructions

“AI Agents — From Language Generation to Task Automation »

M2 TAL Université Paris Cité 2025

Louis Jourdain

Introduction

AI Agents are now being deployed to automate a growing range of complex tasks, from customer support to document generation. These agents are built like *puzzles*: each subtask can be solved by a Large Language Model (LLM), but together they form more powerful, autonomous systems.

One of the greatest challenges of *agentic AI* today, as with NLP in general, is **domain adaptation**. Even the largest LLMs perform poorly on very specific or expert-level tasks, and well-formatted domain data is often too scarce for fine-tuning to be practical. In such cases, **AI agents and innovative agentic architectures** can surpass specialized models, when designed to reason, search, and use external knowledge effectively.

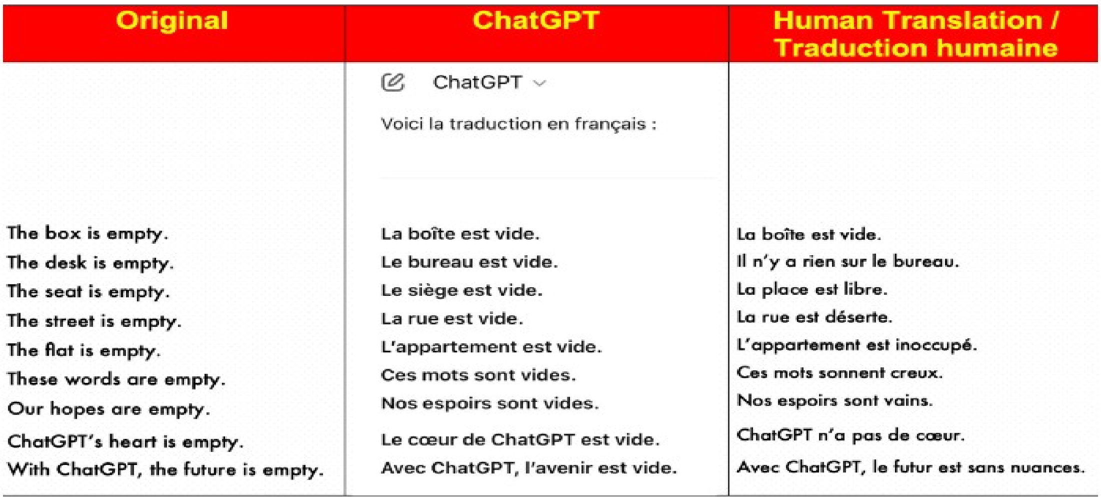
Your mission: **explore how the NLP techniques you know, combined with agent architectures, can solve complex problems that even GPT-5 (or at least the 8B models we’ll use) still struggle with.**

The project will be completed in **teams of 2 or 3 students**.
You may choose between two main topics, and then specialize your subject.

Topic 1 — Agentic Translation

Before 2023, the state of the art in machine translation relied on training a dedicated transformer model for each language pair. This required *large aligned corpora* (I hope you remember Guillaume Wizniewski's class, otherwise I'll tell him…).
Such corpora exist for the ten most spoken languages, but are much harder to find for others, leaving many languages underserved. Projects like Meta's *"No Language Left Behind"* attempted to bridge the gap but with limited success.

LLMs with zero shot prompting have been able to help reach SoTA for the most common pairs of languages. But reaching a good score for a particular metric doesn't mean the task is done yet. LLM can often lack nuance in their translation. [1]



Original	ChatGPT	Human Translation / Traduction humaine
	ChatGPT Voici la traduction en français :	
The box is empty.	La boîte est vide.	La boîte est vide.
The desk is empty.	Le bureau est vide.	Il n'y a rien sur le bureau.
The seat is empty.	Le siège est vide.	La place est libre.
The street is empty.	La rue est vide.	La rue est déserte.
The flat is empty.	L'appartement est vide.	L'appartement est inoccupé.
These words are empty.	Ces mots sont vides.	Ces mots sonnent creux.
Our hopes are empty.	Nos espoirs sont vides.	Nos espoirs sont vains.
ChatGPT's heart is empty.	Le cœur de ChatGPT est vide.	ChatGPT n'a pas de cœur.
With ChatGPT, the future is empty.	Avec ChatGPT, l'avenir est vide.	Avec ChatGPT, le futur est sans nuances.

The main challenges are improving the translation of the « long tail », that it to say the expressions that are the least common in a language. https://arxiv.org/pdf/2506.14702

Recent work has shifted toward **fine-tuning multilingual LLMs** that already possess some *emergent translation abilities*, sometimes using the same limited parallel data as before. This will be the main findings at EMNLP this year (https://www2.statmt.org/wmt25/). Some Prompting techniques have been explored to improve LLM Baseline (https://arxiv.org/pdf/2503.04554)
LLMs are also being used for **data augmentation**, generating synthetic training data (bootstrapping). (https://arxiv.org/pdf/2508.08680)

That's why your first possible topic is:

Agentic Translation for Low-Resource Languages

Human translators rely not only on exposure but also on *meta-linguistic reasoning*: they consult grammars, dictionaries, and contextual clues. They also *know when they're unsure*, and seek external resources to improve their work. We can mimic this by connecting an LLM with **specialized linguistic resources**. This is why using Agents for translation might open new opportunities as it was framed, but not yet well explorer :

https://arxiv.org/pdf/2505.14848 https://arxiv.org/pdf/2504.12891

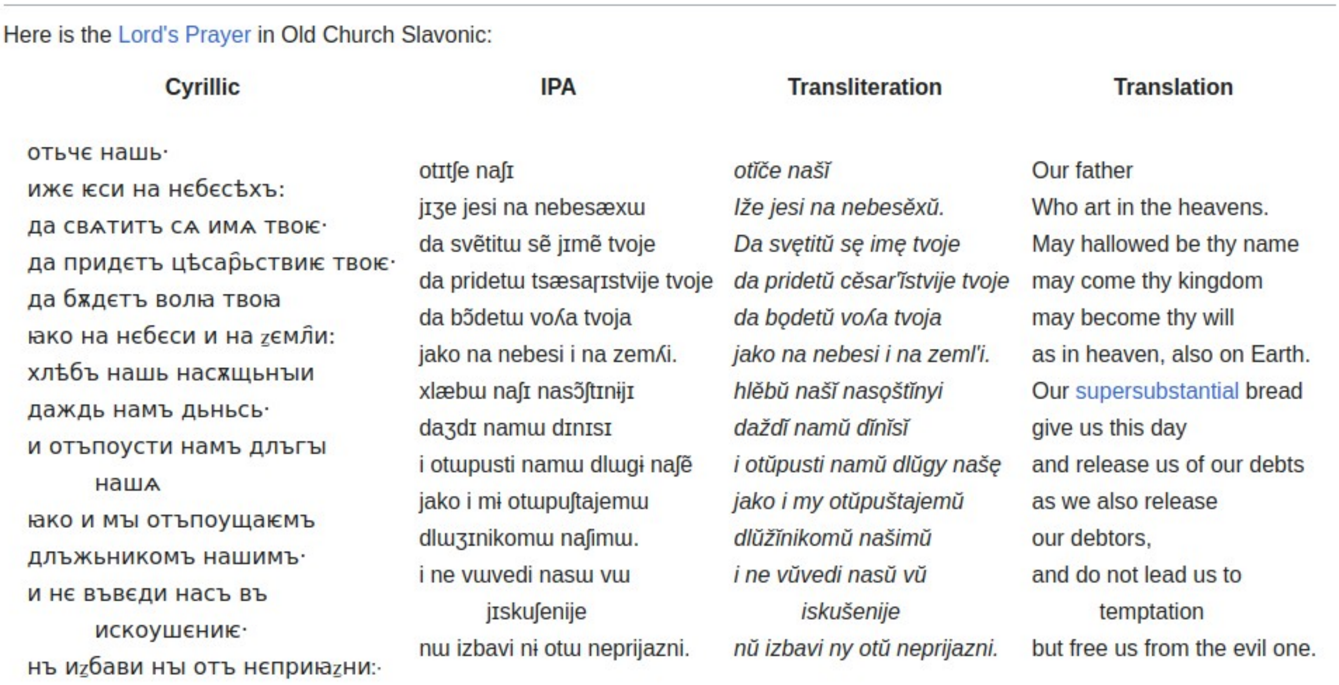
[1] https://www.linkedin.com/pulse/ia-et-traduction-une-exp%25C3%25A9rience-bluffante-antoine-guillemain-qfiae/?trackingId=A5qnOZoD1L03f%2Fdt5B7DDg%3D%3D

To better evaluate the benefit of agentic approaches, you may choose a low resource language:

- A **dead language** I know (Ancient Greek, Old Church Slavonic, Sanskrit), or
- A **Creole language** based on French or English (Haitian Creole, Jamaican Patois, Louisiana Creole, etc.).

Many corpora are available. For Creole languages, see WMT 2025. Part of the mission is gathering the relevant linguistic resources.

You also might face some interesting encoding issues, see for instance this lovely text in Old Church Slavonic :



Here is the Lord's Prayer in Old Church Slavonic:

Cyrillic	IPA	Transliteration	Translation
отьчє нашь·	otɪtʃe naʃɪ	*otĭče našĭ*	Our father
ижє ѥси на нєбєсѣхъ:	jɪʒe jesi na nebesæxɯ	*Iže jesi na nebesěxŭ.*	Who art in the heavens.
да свѧтитъ сѧ имѧ твоѥ·	da svētitɯ sẽ jɪmẽ tvoje	*Da svętitŭ sę imę tvoje*	May hallowed be thy name
да придєтъ цѣсар̂ьствиѥ твоѥ·	da pridetɯ tsæsaɾɪstvije tvoje	*da pridetŭ cěsar′ĭstvije tvoje*	may come thy kingdom
да бѫдєтъ волꙗ твоꙗ	da bɔ̃detɯ voʎa tvoja	*da bǫdetŭ voʎa tvoja*	may become thy will
ꙗко на нєбєси и на ꙁємл̂и:	jako na nebesi i na zemʎi.	*jako na nebesi i na zeml′i.*	as in heaven, also on Earth.
хлѣбъ нашь насѫщьныи	xlæbɯ naʃɪ nasɔ̃ʃtɪnɨjɪ	*hlěbŭ našĭ nasǫštĭnyi*	Our supersubstantial bread
даждь намъ дьньсь·	daʒdɪ namɯ dɪnɪsɪ	*daždĭ namŭ dĭnĭsĭ*	give us this day
и отъпоусти намъ длъгы наша	i otɯpusti namɯ dlɯgɨ naʃẽ	*i otŭpusti namŭ dlŭgy našę*	and release us of our debts
ꙗко и мы отъпоущаѥмъ	jako i mɨ otɯpuʃtajemɯ	*jako i my otŭpuštajemŭ*	as we also release
длъжьникомъ нашимъ·	dlɯʒɪnikomɯ naʃimɯ.	*dlŭžĭnikomŭ našimŭ*	our debtors,
и нє въвєди насъ въ искоушєниѥ·	i ne vɯvedi nasɯ vɯ jɪskuʃenije	*i ne vŭvedi nasŭ vŭ iskušenije*	and do not lead us to temptation
нъ иꙁбави ны отъ нєприꙗꙁни:·	nɯ izbavi nɨ otɯ neprijazni.	*nŭ izbavi ny otŭ neprijazni.*	but free us from the evil one.

Topic 2 — “Llama 3 / Qwen 3 passe l’Agrégation”

Last year, some highschool teachers try to have Chat GPT answer a Baccalauréat (A Level) philosophy exam with poor result : https://www.ekole.fr/blog/on-plafonne-fait-passer-bac-philo-ia-chatgpt-chat-voici-quen-pense-prof. This failure is unsurprising because the model was probably (no detail available) used in a zero shot setting and whithout much guidance leading to an answer much shorter that what was expected. LLMs do not yet have the autonomy alone to automate a 4h task. But maybe an agentic system would have performed much better on the task.
To better challenge agent capabilities for domain adaptation, lets choose an even harder task !

The *agrégation* is a prestigious French competitive exam that qualifies teachers for high school or university. Each year, the Ministry of Education publishes a **specific program** ,texts, periods, or topics, that candidates must master for intense 7-hour written and oral exams. See for instance the program for next year : https://www.devenirenseignant.gouv.fr/les-programmes-des-concours-d-enseignants-du-second-degre-de-la-session-2026-1496

Preparing this exam requires not only memorizing precise knowledge, but also *applying it within strict academic formats* (e.g. dissertation, stylistic commentary, linguistic analysis).

This makes it an ideal playground to explore the limits of agentic reasoning systems.

Even if some people of the field seem reluctant … https://littpo.fr/2025/08/06/passer-lagregation-de-lettres-avec-lia-ridicule/

You may focus on one of the following *épreuves*:

Lettres modernes

- *Ancien français*: phonetic, syntactic, and morphological evolution from Latin to modern French.
- *Stylistique*: stylistic analysis of a literary text following a particular method.
- *Dissertation sur programme*: essay on one of the assigned works.

Histoire

- *Dissertation sur programme*: essay on a historical topic or theme.

Mathématiques

- *Algèbre*: problem solving or structured proof generation.

To start working on one of these topics, you first need to understand in depth what is expected by the student. In the so called « rapports de jury » , « jury reports », available online, https://www.devenirenseignant.gouv.fr/sujets-et-rapports-des-jurys-agregation-2025-1435

you ll be able to find advices for the candidates (useful for prompting?), corrections, reading suggestions that could help you during your development process. Since all this documentation is mostly available in french, I would advice at least one french student in the groups picking this (fun!) topic.

Deliverables Expected

1. Evaluation (Baseline)

- Define a task-specific evaluation dataset and evaluation criteria (can be qualitative).
- Evaluate both a **SoTA LLM (GPT-5)** and a **smaller model (Qwen or Llama 3)** on the same task.
- Analyze weaknesses and identify which subtasks need agentic decomposition.

2. Resource Collection & Tool Design

- Gather linguistic, historical, or mathematical resources.
- Use them to build **tools** (retrieval APIs, lexicons, rule modules, etc.) accessible to your agent.

3. Agent Implementation

- Experiment with different prompting and architectural patterns:
  from simple **ReAct** setups to **multi-agent systems**.
- Evaluate your agent **against the baseline**.

4. A Group Report (around 10 pages)

- Tracking your results over the different iterations
- Explaining and justifying the design choices you did when creating your agents
- Qualitative analysis of the data.

5. A Conference level paper (6-10 pages) presenting your work

- Follow the traditional template :

  1) SoTA and presentation of the problem

  2) Presentation of the architecture of your agent-building

  3) Evaluation and comparison to some baseline / the previous SoTA

  4) Discussion and area for future work.

- You should follow the pagination style expected for a conference paper.

Grading Criteria

Code — /10

- Follows project requirements
- Uses concepts from the course
- Includes runnable code and clear documentation
- Provides an easy-to-use interface (GUI or web)
- Includes observability (logs, traces of runs)
- Tools deployed via API or MCP
- Clean, well-structured code
- Clear run instructions (Docker or script; must work on my machine!)

Report & Article — /10

- Understanding of the problem
- Quality of data analysis and experimental design
- Sound evaluation dataset
- Clarity and coherence of the article, good structure
- Academic strandards in the redaction

I’m thrilled to see you embark on this *agentic adventure*!
Don’t be discouraged if your results aren’t perfect, this project is exploratory by design.
The goal is for you to *understand, in practice,* how to build genuine AI Agents, not the trivial “calendar bots” copy-pasted by influencers, but systems that can tackle *complex, specialized problems* and bring real value to professionals.

And if your results are convincing… we might just polish them into a paper for **TALN later this year**