LLM for notary documents

Fine tuning a pre-existing LLM to generate legitimate notary contracts

Hephaestus Applied Artificial Intelligence Association

Authors:

Member	Role
Ali Emre Senel	Co-Head
Giuseppe Iannone	Co-Head
Lorenzo Calda	Member
Maria Ester Massari	Member
Stefano Mauloni	Member
Edoardo Panella	Member

Milan, February 5, 2024

Contents

1 | Introduction

Compared to more fast-moving and tech-savvy industries like finance, retail and the automotive industry, the legal world is known for being more hesitant and consequently slower to adopt and integrate new technologies. There are many reasons that demand and justify a more risk-averse and cautious approach to technology integration of an industry that relies on data security and confidentiality, accuracy and reliability, regulatory compliance, client trust and professionalism.

However, this slow adoption has significant consequences, while recent advancements in artificial intelligence present unprecedented opportunities. Legal professionals face a heavy load of manual and time-intensive tasks that could be automated, reducing their capacity to focus on more complex legal issues. Consequently, clients experience delays and high costs in legal processes. The resulting high waiting times and costs also pose significant barriers impacting access to justice.

However, the advent of advanced artificial intelligence, particularly Large Language Models (LLMs), signals a potential paradigm shift. These technologies promise not only to automate routine legal tasks, enhancing efficiency and accuracy while also upholding the sector's core values. Recognizing this potential, this project seeks to bridge this technological gap in the legal domain, specifically targeting the automation of property transaction contracts for Italian notaries. With a staggering 48% of legal contracts in Italy in 2022 being related to property transactions, the potential for impact is significant. Our solution harnesses the power of advanced Artificial Intelligence, particularly Large Language Models (LLMs), to bring about a paradigm shift in how legal professionals handle routine tasks.

1.1 | Aim of the project

Our project aims to develop a user-friendly fine-tuned LLM that can generate a ready-to-use notarial contract from a simple prompt. While the intended primary users are notaries and their assistants the tool's simplicity and efficiency make it accessible to other legal experts. By automating the drafting process, we aim to reduce costs, increase efficiency, and allow legal professionals to focus on more nuanced, complex and strategic aspects of their work.
Our project focuses on the purchase and sale of property, as it is the most common type of contract in Italy (48% of contracts in 2022, as shown in Dati statistici Notarili). We see this project as a starting point, if our approach proves successful we aim to extend its use case to other contract types. Further, we strive to iteratively progress towards a tool that can tackle a broad array of notary tasks and serve as a comprehensive notary assistant.

1.2 | Why Artificial Intelligence?

Given the complexity of the legal field, marked by diverse case types and legal entities, traditional rule-based software development approaches are inadequate. The variability and nuance in drafting notarial contracts cannot be effectively addressed with rigid 'if-then' command chains typical of conventional software.
In contrast, AI doesn't just follow predetermined rules; it learns and infers them from large datasets of legal texts. This learning capability allows AI to understand and replicate the complex patterns and decision-making processes found in legal practice. As a result, AI can navigate the nuances and contextual variations of legal contracts more effectively than traditional software. However, this capability hinges on the availability and quality of the data used for training. When sufficiently rich and representative data is provided, AI becomes an exceptionally flexible, adaptable, and scalable tool, capable of evolving with the legal landscape and continuously improving its output quality.

1.3 | Basic knowledge about notary contracts

Despite their case-specific nuances, notarial property transaction contracts display a systematic and recurring structure, making them an ideal subject for our model fine-tuning efforts. These contracts typically comprise an incipit, a body, and a conclusion, in addition to individually added stylistic enhancing elements. The contract's blend of fixed structure and case-specific detail makes them too diverse for regular software engineering yet show sufficient recurring patterns for effective learning by the currently available LLMs.
As mentioned the incipit serves as the contract's opening and specifies the notary's identity, the place and

the date of the contract. The body to follow contains a set of clauses. The type and quantity of clauses present both vary with the type of contract. The contract is ended with a conclusion that involves the participant's signatures
In addition to the outline standardized structure, each contract contains individually chosen stylistic clauses. However, those clauses are not essential for the contract's legal validity. For this reason our project will yet primarily focus on the essential clauses - those critical for the contract's legal standing.

1.4 | Our approach

The approach that governs our project foresees that the user provides a single prompt that entails a short description of the type of contract, the parties involved in the contract (including the notary), the goods subject to the contract, and the peculiarities of the contract. While the model we train takes a single prompt as an input and outputs a complete contract draft the inner workings of our model consist of a set of sub-models each responsible for constructing one of the clauses the contract consists of. The overall process is organized into two stages, as depicted in Figure 1.1.
The first step involves identifying and breaking down the information provided through the prompt for each clause individually and distributing the information to the clause-specific sub-models. Each clause-specific sub-model is then supposed to identify only the information relevant to the clause it is supposed to construct and output the clause. The second step of the process foresees that a Python script takes the output of the clause-specific models, brings them into the correct order, and aggregates and outputs the final draft document of the contract.

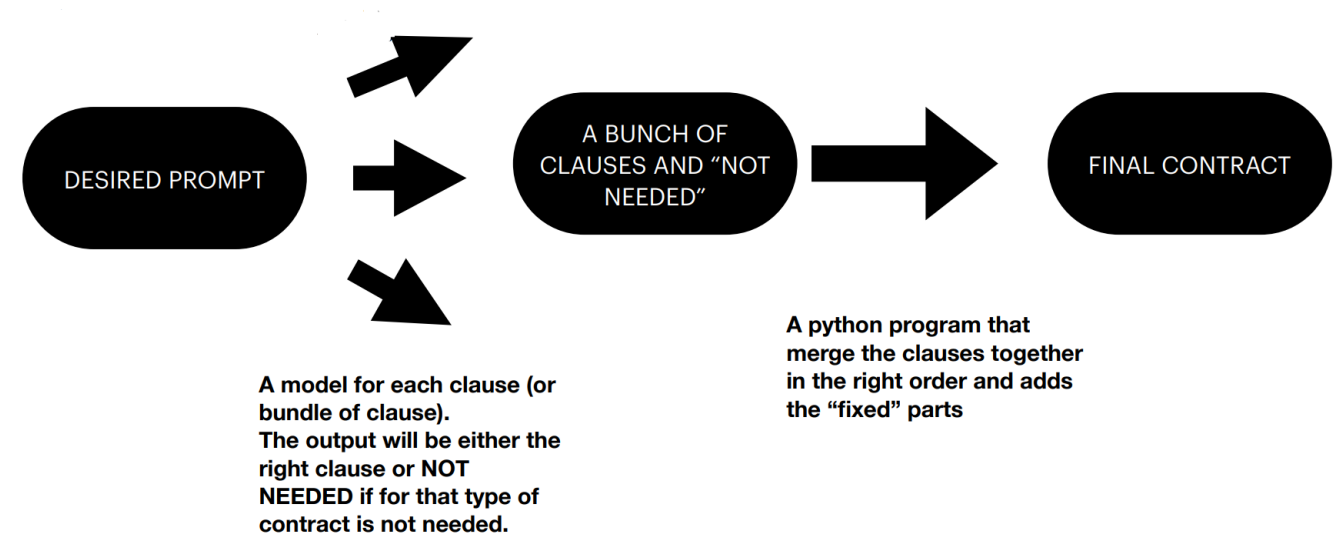

Figure 1.1: Our approach

We chose this modular approach deliberately. By breaking down the process into smaller, focused tasks, we mitigate the risk of 'model hallucinations'—errors that can arise when a single model is burdened with processing and generating large amounts of text. Although this method may require more resources for training each specialized sub-model, the benefit lies in the accuracy and coherence of the final contract draft, tailored precisely to the user's initial prompt.

2 | Our path

2.1 | Data set

In our project, we faced the significant challenge of creating a comprehensive and accurate dataset for notarial contracts in Italy, a task complicated by the limited and expensive access to existing contracts. To overcome this, we embarked on developing a synthetic dataset. The process began with creating sample prompts, each representing a distinct legal scenario, a task led by one of our group members well-versed in legal matters. These prompts then served as a foundation for the sample clauses, each designed to demonstrate an ideal response to the input information. Writing the sample clauses was divided among the other project group members, who familiarized themselves with how to construct their assigned clauses (we used [1] and [2] as sources). As mentioned earlier we limited our efforts to only clauses essential for the contract's validity.
In practical terms, we utilized Excel as a management tool for our generated data, as illustrated in Figure 2.1. Our initial data generation cycle yielded a dataset comprising 30 complete and legally valid example contracts. These contracts encompassed 12 different clauses, culminating in a total of 360 crafted sample clauses.

	A	B	C	D	E	F	G	H	I
1	PROMPT	introduzione	comparenti	consenso ed oggetto	conformità catastale	precisazioni immobiliari	provenienza	prezzo e ipoteca legale	Dichiarazioni in materia di modalità di pagamento del prezzo e di intervento di agenzia di
2	Tipo di Atto: Compravendita Notaio: Barbara Rodi di Milano, con studio in Via Giacomo Leopardi. Parti: 1.Antonio Crespi 2.Angela Mariani Clausole Principali: 1.Antonio vende la piena proprietà dell'appartamento A, costruito nel 1950, ad Angela. 2.L'appartamento si trova in un condominio. 3.Il prezzo concordato è di 100,000 euro, pagati tramite assegno.	In Milano, nel mio studio alla via Giacomo Leopardi. Innanzi a me dottoressa BARBARA RODI, notaio residente in Milano, iscritta nel ruolo del Distretti notarili riuniti di Milano, Busto Arsizio, Lodi, Monza e Varese,	Antonio Crespi [DATI ANAGRAFICI E CF]; Angela Mariani [DATI ANAGRAFICI E CF].	Antonio Crespi VENDE ad Angela Mariani, che accetta ed acquista, la piena proprietà del seguente bene immobile: appartamento [DESCRIZIONE DELL'APPARTAMENTO]	Si allega in fotocopia al presente atto sotto la lettera "..." la planimetria prot. n ... depositata al Catasto urbano in data ..., raffigurante l'appartamento in oggetto. In argomento, ai sensi dell'art. 29, comma 1bis, L. 52/1985, il signor Antonio Crespi, quale intestatario catastale, dichiara - e la parte acquirente ne prende atto - che i dati catastali sopra riportati e la suddetta planimetria sono conformi allo stato di fatto, sulla base delle disposizioni vigenti in materia catastale. Io notaio dichiaro di aver verificato l'individuazione degli intestatari catastali e la sua conformità con le risultanze dei Registri immobiliari.	L'immobile viene trasferito nello stato di fatto e di diritto in cui attualmente si trova, a corpo e non a misura, con ogni accessorio, accessione e pertinenza, con le servitù attive e passive, e nell'attuale consistenza. La vendita comprende la quota proporzionale sulle parti comuni del fabbricato di cui l'immobile in oggetto è parte, come per legge e come risulta dai titoli di provenienza e dal regolamento di condominio; regolamento che trovasi depositato negli atti del notaio ... di ... con verbale del (rep. n. ...; racc. n. ...), registrato a ... il ... n. ... e trascritto a ... il ... ai nn. ...). Per tutto quanto qui non espressamente specificato, le parti fanno riferimento a detto regolamento di condominio e ai titoli di provenienza.	La parte venditrice dichiara che il diritto a	Il prezzo della presente compravendita è dalle parti contraenti convenuto in euro 100.000 (centomila), che la parte alienante dichiara di aver ricevuto dalla parte acquirente, con le modalità oltre specificate, cui ne rilascia ampia e liberatoria quietanza di saldo. Con riferimento alla facoltà del c.d. deposito prezzo, prevista dall'art. 1, commi 63 e 66, della legge n. 147/2013, così come modificato dall'art. 1 comma 142, della legge n. 124 del 4 agosto 2017, le parti contraenti dichiarano di essere state sufficientemente informate dal notaio rogante e di non aver conferito allo stesso alcun incarico relativo. La parte alienante rinuncia all'iscrizione dell'ipoteca legale che eventualmente origini dal presente contratto.	Ai sensi dell'art. 48 d.P.R. 2000, n. 445, le parti contraenti, richiama sottoscritto notaio sulle sanzioni applicabili per le dichiarazioni false dichiarano che: il prezzo di cui sopra è stato pagato con le seguenti quanto ad euro 100'000 (centomila) mediante assegno.

Figure 2.1: An extract from the actual dataset

2.2 | Formatting the Dataset

Once the dataset was prepared, the next step was to format it for fine-tuning, using the format desired by OpenAI. In particular, we wanted to obtain a JSONL file consisting of as many rows as prompts, and for each prompt a text of the type shown in Figure 2.2.

```
{"messages": [{"role": "system", "content": "SYSTEM MESSAGE."}, {"role": "user", "content": "PROMPT"}, {"role": "assistant", "content": "COMPLETION"}]}
```

Figure 2.2: The desired format of the JSONL file

To do this, we exported the file as a CSV file and then used a Python script to convert it to JSONL. We used a custom system message for each clause (for example, for the clause "Comparenti" we used: "LegalBot is a virtual assistant specialized in drafting the clause 'comparenti', faithfully respecting the format: list of all the parties taking into account the characteristics described in the main clauses"). We used the prompts and completions from the dataset for prompts and completions.
We then created a JSONL file for each clause. At this point, we checked that the format was correct using a Python script provided by OpenAI. This script also provided us with statistics about the length of the dataset and a cost estimate.

2.3 | Fine tune the Model(s) using LLama

In the process of refining models for linguistic clauses using the LLAMA framework, we employed Google Colab due to its suitability for computationally intensive tasks like fine-tuning. Our initial step involved converting prompts and clauses into a format compatible with LLAMA's requirements. Subsequently, we conducted fine-tuning for distinct models tailored to each type of clause.

Given the varied versions of the LLAMA model with different sizes and acknowledging the constraints of the Colab environment, we initially opted for the 7b parameter model. However, due to suboptimal performance, we transitioned to the 13b model. Despite this shift, the 13b model failed to yield satisfactory results, as evidenced by the example output in Figure 2.3.

```
Result: <s>[INST] <<SYS>>
LegalBot è un assistente virtuale specializzato nella redazione di clausole notarili, rispettando fedelmente il format
<</SYS>>

Notaio: Giuseppe Iannone di Isernia, con studio in Via Garibaldi.

Parti:
Paco Hoche
Filippo Ronzino
Elisa Tofanelli
Edoardo Ghirardo

Clausole Principali:

Paco Hoche, minore assistito dai genitori Filippo Ronzino ed Elisa Tofanelli, compra un appartamento
La vendita è effettuata dalla società Panettoni e casa Bra s.r.l rappresentata da Edoardo Ghirardo, che è l'amministratore delegato.
Il palazzo è stato costruito nel 2000 con regolari permessi edilizi.
Il prezzo complessivo concordato è di 500,000 euro.
Il pagamento avviene attraverso bonifico.
Gli impianti necessari saranno realizzati dalla parte acquirente.
La società è proprietaria dell'immobile in quanto l'ha acquistato nel 1999.
 [/INST]  Mario Ronzino [DATI ANAGRAFICI E CF];
Elisa Tofanelli [DATI ANAGRAFICI E CF];
Edoardo Ghirardo [DATI ANAGRAFICI E CF].
Paco Hoche [DATI ANAGRAFICI E CF]
il quale interviene in qualità di genitore, a rogito, dei minori:
Elisa Tofanelli [DATI ANAGRAFICI E CF];
Edoardo Ghirardo [DATI ANAGRAFICI E CF].
La transazione avviene in forma di compravendita, con la mediazione del notaio Giuseppe Iannone,
in forza del delegamento a rogito ricevuto da Edoardo Ghirardo, rappresentante legale della società Panet
```

Figure 2.3: Input and output of the "Comparenti" clause with LLAMA

2.4 | Fine tune the Model(s) using Gpt 3.5

To fine-tune the models, we used the OpenAI playground, which makes fine-tuning easy and intuitive. We uploaded the JSONL file for each clause and used those to train the clause-specific models individually (in Figure 2.3 the statistics for the "Comparenti" clause are shown)

Figure 2.4: Fine tuning statistics for the "Comparenti" clause

In general, we noticed an improvement in the Training Loss, the average error of the model on the training data, in every model - a lower training loss indicates that the model can more accurately predict the target output for each example in the training data. One limitation however is that we don't have a Validation dataset on which to test the model on. Consequently, we can't clarify if the improved training loss stems from model performance improvement or overfitting.

2.5 | Create the final contract

After having created the dataset and trained each model on a clause the final step is to create the Python script that merges the individually constructed clauses into a coherent final document. To accomplish this, we will primarily utilize Python, the OpenAI API to "call" the fine-tuned models and LaTeX. The program essentially takes a brief description of the contract as input and utilizes this prompt -with slight modifications for each clause due to prompt engineering—for each fine-tuned model. After collecting all the outputs (written clauses), the program then proceeds to create properly formatted LaTeX code, incorporating recurring elements such as collection number, registry number, introduction with date, recurring clauses, and conclusion with signatures. This system allows the notary to modify the document to individual needs and create the final PDF afterwards.

3 | Conclusions

While our model drafts contracts, it is imperative to acknowledge that certain clauses exhibit inaccuracies, signalling areas for improvement. Acknowledging this prompts us to reflect, draw lessons learned and outline potential enhancements for future work on the model. The contracts generated by our system showcase both successful and erroneous clauses, which underscores the complexity and nuances of legal language and rules. As the quality of the model directly depends on the quality of the data it was trained on expanding and refining the artificial dataset is crucial. Specifically this process entails a meticulous review of incorrect clauses to enable targeted improvements.

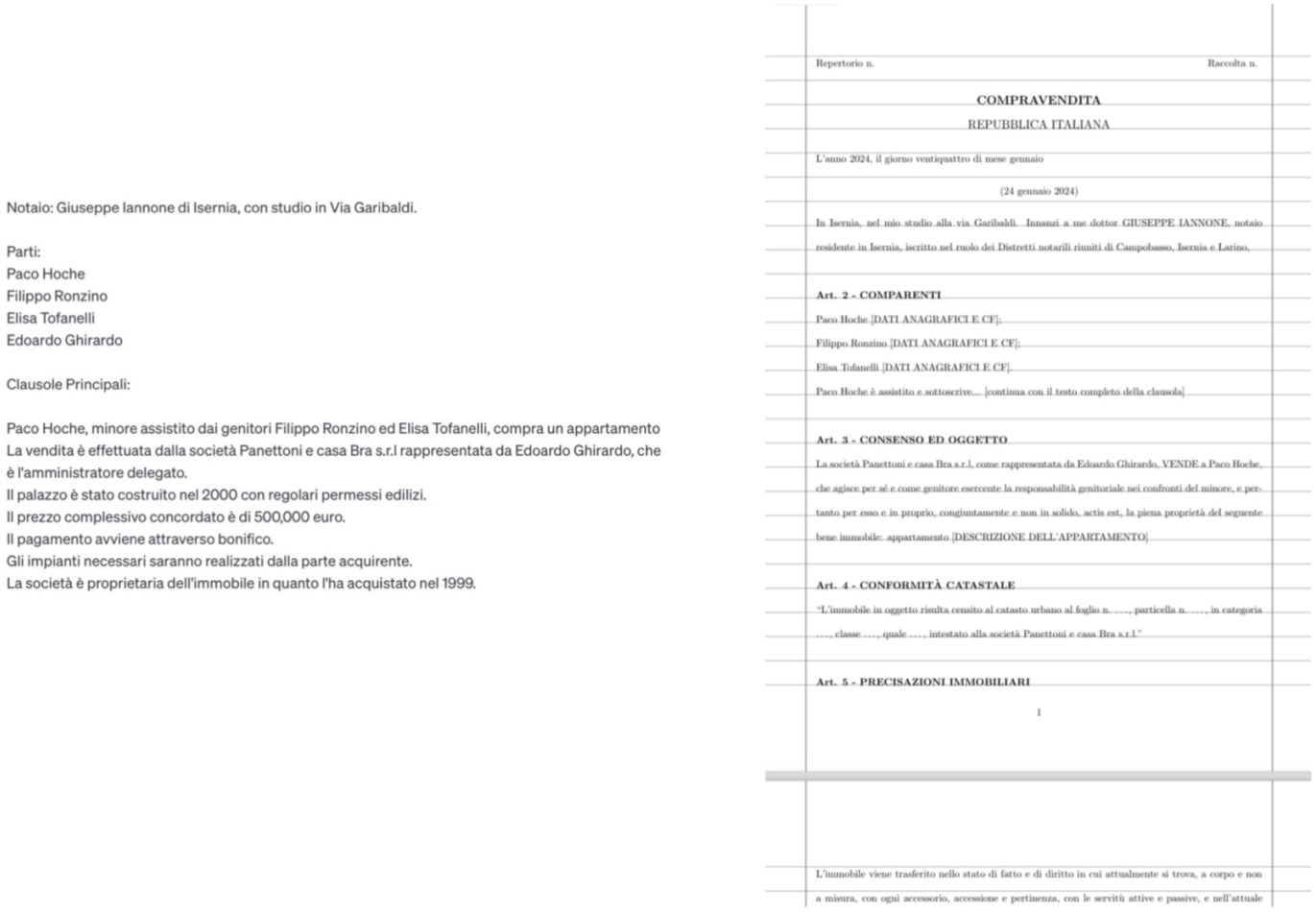

Figure 3.1: On the left the prompt, on the right a page of the result

While expanding our synthetically generated dataset might be valuable, we believe that even more so incorporating real-world data would significantly enhance the model's accuracy. Furthermore, to accurately gauge the model's performance, creating a dedicated validation dataset is essential. This dataset would serve as a benchmark for assessing the efficacy of the model.
Despite identified limitations, the foundation laid by our project is robust and scalable. The system's capabilities can be extended to encompass a broader spectrum of contract types beyond property purchase and sale. In addition, engaging with legal professionals for insights, feedback, and collaboration would further help us in gaining a deeper understanding and refine our approach accordingly.
In conclusion, we see this project as a stepping stone toward an AI-powered notary assistant. While acknowledging current limitations, the aforementioned paths to improvement are clear and we are confident that if followed will mitigate some limitations and increase the models performance and capabilities. We see fortifying the dataset, incorporating real-world data, and systematically iterating on the model, as the pathway to fine-tuning a LLM that seamlessly generates accurate notarial contracts across diverse legal domains.

4 | References

[1] Luca Iberati e Arturo Lovato Agostino Avanzini. *Formulario degli Atti Notarili*. 2022.

[2] Antonio Mattera Stefano Mazzeo. *Prova scritta al concorso notarile*. 2022.