Categorizing e-Invoice Data through Local LLMs

Francesco Ascione
Italian Revenue Agency
Taxpayers Compliance and Enforcement Division
Risk Analysis and Tax Compliance Research Unit
Data Science Office

05/06/2025

Outline

- Legal Basis and Goal
- Workflow
- Local LLM environment
- Target Categories
- Model Instruction
- Results
- Conclusion

Legal Basis and Goal

Challenge

The deductibility of VAT associated with a taxpayer's purchases depends on the *inherence* of these purchases to the taxpayer's business activity.

For evaluating this inherence, textual data analysis on e-invoice descriptions could be very valuable.

…but invoice descriptions are filled with complexities – brand names, abbreviations, technical language. Previous attempts using anomaly detection faced challenges in interpreting these nuances.

Goals

- telling non-inherent purchases from the others (in order to identify taxpayers that have unduly deducted VAT on purchases)
- assigning relevant product or service categories to e-invoices

Legal Basis and Goal

Key features of the solution and constraints

- The solution uses Large Language Models (LLMs) to understand complex invoice descriptions
- The solution must operate locally – no data over the network
- Response times under 5 sec per invoice
- Designed for testing on commonly available hardware with basic requirements.

Focus of this presentation

Today we discuss an experimental approach. We will delve into qualitative insights from a small sample to assess the model's capacity of handling ambiguity.

We will not present a full-scale quantitative analysis (i.e., interpreting results on a large dataset) at this stage.

Workflow

- Extract sample of e-invoice data
- Clean data
- Configure local LLM environment
- Define target categories
- Instruct the model (prompt engineering)
- Test models
  Interpret results

Local LLM environment

Basic requirements. LLMs are optimized for the processing power of GPUs. However, our experiment focuses on their feasibility on common hardware (a standard laptop, 16GB of RAM).

Key factors. Two factors directly impact the model inference speed: model size (number of model parameters, expressed in billion) and model quantization (precision of the model's numerical weights).

Setup. With these considerations in mind, we have implemented a local environment for running LLMs using two popular frameworks: Hugging Face Transformers and Ollama.

Model tested	*Authors*	*Notes*	
gemma-3-1b-it, f32 quantiz.	Google	1B parameter, full precision quantiz.	
gemma3:4b, Q4_K_M quantiz.	Google	4B parameters, less precise quantiz.	CHOSEN MODEL under 4sec for response
llama-3.2-3B-Instruct	Meta	3B parameters, multilingual	
deepSeek-R1-Distill-Qwen-1.5B	Deepseek AI	Reasoning multilingual model	
minerva-7b-instruct-v1.0-q4_0	Sapienza University of Rome	Italian language model	
LlaMantino 2	Università di Bari	Italian language model	

The **temperature parameter** is another useful setting to keep in mind. It controls the randomness of the LLM's output. Lower values make the output more focused and deterministic, while higher values can lead to more creative or diverse responses.

Target Categories

Our next step is to define the specific product or service categories that the model will assign to invoices.

These categories are being designed specifically to assess the inherence of purchases to the taxpayer's business activity.

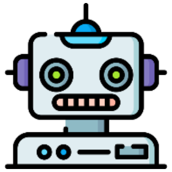
Food & Beverages	Education & Training	Travel
Clothing	Professional Services	Events & Entertainment
Watches & Jewelry	Home & Office	Financial expenses & Investments
Health & Wellness	Vehicles & Transportation	

Under national regulation, these categories are often classified as non-inherent.

For instance, these categories would be non-inherent for legal or financial professionals.

Model Instruction

Providing instruction to the model involves two key components:



System Prompt
This prompt defines the LLM's role, personality, or overall behavior. It contains general directives.

User Prompt
This is the specific question or instruction related to the input data (in our case, the invoice data).

First attempt at communicating the model our goal:

System Prompt: You are an assistant that responds concisely, without giving explanations.

User Prompt: Consider this product: “{description}”, sold by a company involved in {seller’s business activity}. To which of these categories does the product belong? {category list}. Indicate only the category, without explanation.

To improve categorization accuracy, we iteratively refined our prompts. Ultimately, we made the system prompt more specific about the LLM's role and output constraints, and we simplified the user prompt.

System Prompt: You are a commercial assistant who assigns a category to each product sold. The categories you can assign are: {category list}. Never provide explanations for your answers, indicate only the category.

User Prompt: Consider this product: “{description}”, sold by a company involved in {seller’s business activity}.
To which category does it belong?"

Results

Here we present a subsample of the invoices we examined, drawn from a dataset of products and services purchased by lawyers.

INVOICE DESCRIPTION (translated)	SELLER'S CORE BUSINESS ACTIVITY	CORRECT CATEGORY	MODEL RESPONSE
lawyer course in-pesron 2019 Pescara offices - 1st payment	cultural education	education & training	«education & training»
fees // fees for legal research and translation activities carried out in the month of January 2019	research and developm. in social sciences	professional services	«professional services»
third floor apartment renovation with replacement of windows and new flooring (…) - amount agreed before start of the work	construction of other civil engineering works	home & office	«home & office»
5 red prawns 20.00	retail sale of fish in specialized stores	food & beverages	«food & beverages»
sponsorship for competitive activities, advertising for your brand and name	sport activities	financial exp. & investments	«events & entertainment»
earrings met/res/str dore/blanc nacre/cristal uni sans finition	wholesaling of perfumes and cosmetics	watches & jewelry	«watches & jewelry»
swc super smash bros ultimate	retail sale of games and toys in specialized stores	events & entertainment	«videogames»

seller's business activity not helping much here

wrong response

technical language not confusing the model

"inventing" new category, but correct

Conclusion

- The approach we followed is still early-stage, but we showed that local LLMs are promising for handling the ambiguous descriptions often found in invoices.
- LLMs are likely to become even smaller and more powerful in the near future. This trend will allow them to be effectively used by national and trans-national institution as well as their trusted tech partners.
- Possible improvements to the proposed solution:
  - dedicated hardware (GPU)
  - providing the model with a clear definition of the categories (increasing context size)
  - providing the model with invoices already categorized (using RAG?)
  - using LLMs for text disambiguation only, then categorizing invoices with a different ML technique

Thank you for your attention

APPENDIX

Configuring Environment

We have implemented a local environment for running LLMs using two popular frameworks: Hugging Face Transformers and Ollama.

Hugging Face

Method 1: Hugging Face transformers

- Install python
- Install transformers library by Hugging Face

Hugging Face is a leading open-source platform providing a vast hub and tools for pre-trained machine learning models.

- Install deep learning framework (pyTorch or TensorFlow)
- Download the desired pre-trained model and its tokenizer directly from the Hugging Face Hub

Method 2: ollama

- Download ollama from ollama.com and install

Ollama is a user-friendly tool designed to simplify the process of downloading, setting up, and running open-source LLMs locally across various operating systems.

- Pull model from Ollama (or download it from Hugging Face library)
- Run via command line or using the ollama python library