Pydeia

Know Yourself. Find Your Path.

This instrument was born to address a central problem in what appears to be one of eighteeners most troubling questions: which university am I going to choose? Where should I go? On what grounds am I shaping my future career?...

This is where Pydeia comes in to play, an agentic AI especially developed for young women and man who don't really have a clear idea about what they are going to do with all the possibilities in their hands.

From an user friendly vocal interface, built to listen to your needs and propensions, to an in deep profiling that ultimately brings to a short tailored list of facilities that will most likely fit the experience you're looking for!

No more need for counselling, school advisors or open-days, Pydeia is going to be everything you need

The Pipeline

Fully exploiting the Datapizza framework, we were able to orchestrate and automate the full automation process: first of all, we wanted an intuitive GUI that could grant the possibility to make the user feel at ease with the instrument and, focusing the analysis on the attitudinal/psychological side we chose speech as the best mean to vehicolate and transmit information directly from the consumer to the bot; behind the chatting experience, there is a layer of agents and structured output smart functions that organize retrieved natural language information in structured data that is then embedded in vectorial form together with all faculties and associated metadata, to make this process as fluent and seamless as possible, we exploited another important Datapizza tool that allowed us to work with small, local machine learning models, capable of doing an astonishingly efficient and qualitative job when given very specific tasks; every chosen feature of the student and universities was embedded separately and matched via cosine distance search, to ensure precision in the evaluation of the most suitable choices. All distances, were then put together as scores, weighted by a factor, also chosen by an LLM, that reflects the infrared importance of each attribute from the future bachelor student perspective. The overall scores obtained is finally used to compile a top-3 list of the best matching faculty candidates that is presented with a list (generated by a dedicated agent) of pros and cons of such choice, keeping in mind the whole discussion and comparing the three options with each other.

```
"You are a university guidance agent.
You must remember that the context is
only Italy. Do not ask about foreign
countries.
Given a text (spoken or written) by a
high school student,
you must FILL an object with the
following fields (Pydantic model
'Info'):

- academic_profile: what the student
is more inclined / interested to study
  (e.g. "Computer Science,
Engineering" or "NO: Medicine, Law")
- aspiration_values: how motivated /
serious / proactive the student is
  (e.g. "very motivated, disciplined"
or "not very motivated, lazy")
- lifestyle_preferences: desired
lifestyle
  (e.g. "party-oriented, sporty" or
"calm, studious")

- budget: yearly budget in euros if
explicitly stated or reasonably
deducible (otherwise None)
- origin: city of origin (if never
mentioned, leave as None)
- location: Italian city or area where
they would like to study.
```

```
  If the student has NO preference, or
says they are indifferent, or says
that many cities are fine,
  then set location = "" (empty
string).
  Use None ONLY if the text is unclear
and it is not possible to deduce
anything.
- gpa: school average on a 0-10 scale
(if not specified, None)
- max_distance: maximum acceptable
distance in km from the origin city,
if deducible

- far_from_home: if the student wants
to go far from home (live away from
family),
                  or if they prefer to
stay close; if no information is given
in this instance, keep this at None
- english_language: if they are open
to English-taught courses (if stated
explicitly),
                     or if they say
they DO NOT want English; if no
information is given in this instance,
keep this at None
- dorms_nearby: whether they care
about university dorms / residences
- admission_test: whether they are
available to take an admission test,
                   or if they reject
tests; if no information is given in
this instance, keep this at None
-   extracurricular_activities:   if
clubs/extra activities are important
to them,
                                 or   if
they do not care; if no information is
given in this instance, keep this at
None

VERY IMPORTANT:
- Do NOT make up values.
- Leave fields as null/None if the
text does NOT specify them or is not
clear enough.
- You will receive multiple messages
from the same person over time. Each
time, fill in only
  what you can deduce from THAT
specific message."
```

Prompt used to make the structured-output generator fill the attributes of the class-like object created to structure data and update it at every iteration

One of the main focuses and challenges on the input-processing side, was making good use of modern prompt-engineering theory in order to make the large-language-model execute the task we thought as close as possible to how we wanted it to be done. This thorough refining process led to the possibility of obtaining a satisfactory output from a relatively cheap and out-of-date model such as gpt-4o-mini; we here the first escalation possibility of our work, using this same tool, side-to-side with state-of-the-art llms could drastically improve performance and precision, granting the possibility of a more refined profiling process with direct consequences in the output of the pipeline.

"It sounds like you're still in the early stages of deciding on your university options! I'm curious, what kind of lifestyle do you envision for yourself during your university years? Would you prefer a more sociable experience, going out often, or a quieter, more studious environment? How important is this aspect in your final decision?"

This is an example of an early stage question the program is asking to dig deeper in the users habits and expectations from university life; going into this project we really

thought and focused on the possibility to probe the psychology of the user in a non invasive way and early tests already demonstrate the propension of the model for such kind of questions.

Postprocessing of data, as mentioned, goes forth via internal representation of the data that is embedded in large dimensional vectors thank to the coordination with the granite-embedding:30m model, that is more than a thousand times smaller than multimodal models but works perfectly for a (up to now) small dataset, resulting a more sustainable solution both in environmental and economical terms.

Scalability

What we believe to be the beauty of this work is actually the possibility not just to expand it way beyond Italy's borders, but the chance to make it even more competent in reading into emotions and moods of users.
First of all, collecting a much bigger amount of data (faculties) would path the way to the creation of a populated vector store, largely enhancing the precision of vectorial search and thus the quality of the output.
As already stated, another fundamental dimension of improving, would be the usage of a much more powerful LLM for queries and data retrieval, the Datapizza framework makes it very easy to switch from a model to another,

making it possible to effortlessly keep an up-to-date technology working in the back-end of the web page.
ML light-fast improving would also be key in empowering the agent with better and more precise speech-to-text transcription and make it possible to employ other types of interface-student means of communication such as video and handwritten files. Also new tools are emerging for emotion recognition from audio files, instruments that could and will revolutionize the agentic AI landscape and applications.
Adding new characterizing features for a more in-depth profiling is also a challenge that would be necessary to improve the experience and satisfaction of the user.

Earning Opportunities

We believe that, from a project like this could benefit both the to-be bachelor students and the academic institutions; our vision and experience before that, taught us that there is a lot of people that walk in the dark when faced with this potentially life-changing decision and they always get little to no help from peers or teachers, we do think than that there would be a good amount of young students willing to pay a small fee to be helped and guided in this new stage of their lives. In a complementary way, universities could benefit from statistical

insights on the profiles of the students that could be interested in enrolling in one of their faculties and Pydeia could also host the application process for example, making it easier for both parties to handle the know laborious paperwork behind university enrolment. In this way, our framework could really serve as a bridge between the two entities, reshaping the way we think of student-university collaboration.