Contribution of AI to validation

Marco Lippi
marco.lippi@unifi.it

Joint work with Simone Marinai and Valeria Nardoni

UNIVERSITÀ
DEGLI STUDI
FIRENZE

Artificial Intelligence for clinical data

Recent **generative AI** models have reached stunning results in many applications related to **natural language processing** and **understanding**

Clinical data is not an exception, and there is a **growing interest** in exploiting Large Language Models (LLMs) for the **automatic analysis** and processing of textual documents related to healthcare

- Information revtrieval
- Text classification
- Question answering
- Document summarization
- ...

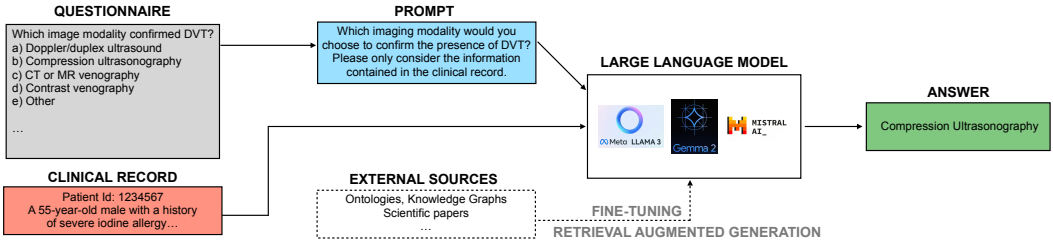
UNIVERSITÀ DEGLI STUDI FIRENZE

AI for the SeValid project

Can AI work alongside humans in **complex**, **time-consuming** and **prone-to-error** tasks related to clinical data, such as questionnaire filling?

- Can we **compare** humans and AI?
- Can we **integrate** humans and AI?
- Can we **substitute** humans with AI?

Questionnaire filling from clinical records



Why LLMs?

UNIVERSITÀ DEGLI STUDI FIRENZE

Questionnaire filling from clinical records

Why LLMs?

- No need to build a labeled training **corpus**
- No need to build a **classifier** for each and every question
- Possibility to exploit **background knowledge**
- Possibility to exploit **zero-shot** or **few-shot** learning
- Possibility to exploit **reasoning** via Chain-of-Thought

UNIVERSITÀ DEGLI STUDI FIRENZE

Prompt engineering (1/2)

```
Read the patient's medical record and respond exclusively with 'yes', 'no', or
'unknown' for each of the following cardiac symptoms.  Do not make inferences,
assumptions, or provide additional information.  Each symptom must be assessed
**independently** and based strictly on what is explicitly stated in the patient's
medical record.  If a symptom is not explicitly mentioned, respond with 'unknown',
even if related terms or conditions are described.

Cardiac symptoms:
- Acute chest pain or pressure
- Palpitation
- Dyspnea (at rest, with exercise or lying down)
- Diaphoresis
- Sudden death
- None of the above
- Unknown if any of the above are present or absent
```

Prompt engineering (2/2)

```
**Important Notes:**
1.  Do not assume that similar terms or related conditions (e.g., arrhythmia and
palpitation or retrosternal pain and acute chest pain) are equivalent unless
explicitly stated.
2.  Do not assume sudden death be no if is not explicitly stated is uknown.
3.  If no information is provided about a symptom, respond with 'unknown'.
4.  Put yes or no only if the words put between the options are present in the
medical record; if not it is "unknown"

Patient record:  BLA BLA BLA...
```

UNIVERSITÀ DEGLI STUDI FIRENZE

Experimental results

We tested our LLM on 38 real cases (in Italian!) collected via the SeValid protocol

- The LLM provided answers to all the questions in the questionnaire
- Answers were collected and processed, to extract levels of risk
- We measured concordance on **single questions** (computer science perspective)
- We measured concordance on **levels of risk** (clinical and epidemiological perspective)

UNIVERSITÀ DEGLI STUDI FIRENZE

Experimental results

Results: the **computer science** perspective

- Over 80% agreement on average for each single question
- Some questions are more difficult than others
- Prompts can be refined and further improved
- More recent and larger models could be used
- Consider changing the language of the prompt (or translate the record)

UNIVERSITÀ DEGLI STUDI FIRENZE

Experimental results

Results: the **clinical and epidemiological** perspective

- The LLM has been much more cautious in providing answers
- Thus, often the questionnaire filled by the LLM produces a **non-assessment**
- 30 patients: CASE both for LLM and human
- 5 patients: CASE for human, NA for LLM
- 2 patients: CASE for human, NO-CASE for LLM
- 1 patient: NA for human, CASE for LLM

UNIVERSITÀ DEGLI STUDI FIRENZE

Some considerations for final discussion

- Is it safer to have the LLM stick with the information in the clinical record?
- Would it be suitable to **run multiple models** in order to mitigate variance?
- Would it be suitable to **run different models** for different clinical cases?
- Would it be reasonable to **know in advance** whether the patient belongs to the set of narrow or possible patients, or not?