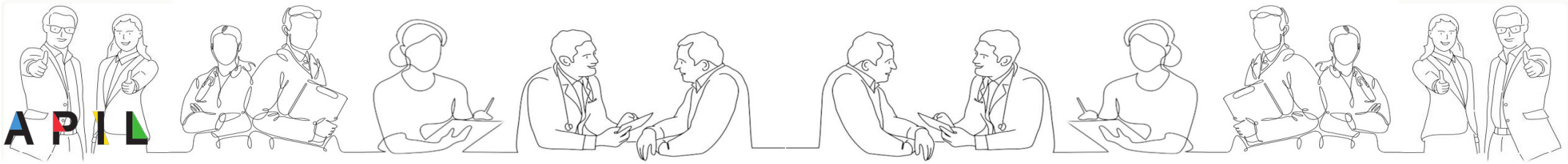
# NLP, Ethics & Standards
# A point of view from French companies

Hugues de Mazancourt, APIL, Datapolitics

Alain Couillault, APIL

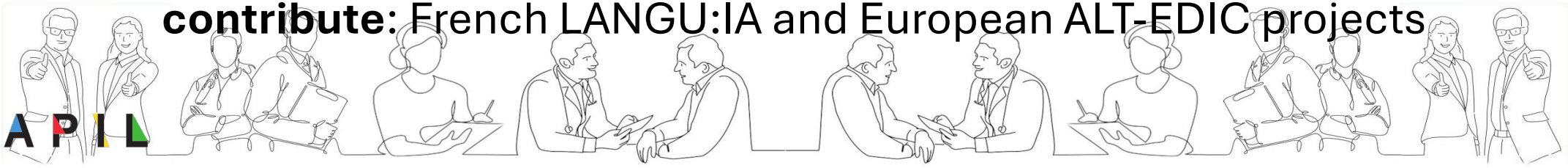


# Who we are

APIL: our members are French companies providing services or software solutions for/with Natural Language Processing

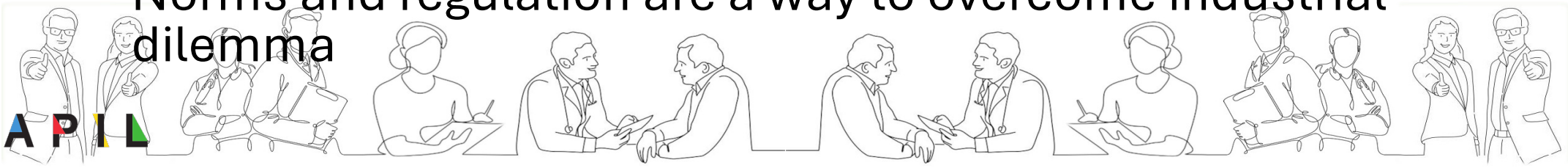
Our actions:

- **assemble**: meet-ups, dedicated days, involving industrial and academic bodies
- **collaborate**: with our partners : Le VoiceLab, HubFrance IA, ATALA (Jeunes TALents)
- **disseminate**: through our website www.demotal.fr
- **contribute**: French LANGU:IA and European ALT-EDIC projects



# The upsurge of LLMs in industry

- For an industrial company in the NLP field, it is today impossible to sell products (or raise funding) if the term “LLM” is not present (at least) in the product description
- This raises issues linked to ethics:
  - Everlasting questions still not answered
  - New customer requirements
  - New technical issues
  - New regulations
- Norms and regulation are a way to overcome industrial dilemma



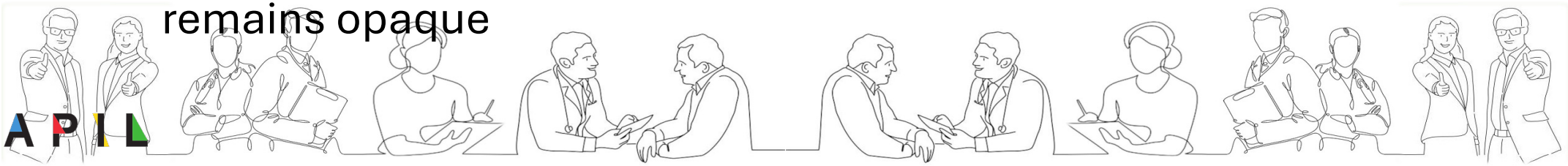
# Training datasets

Where do the data come from?

- **raw** vs **enhanced** vs **synthetic** data
- copyright issues, privacy/GDPR compliance, etc.
- ethical issues on data collection/annotation (cf. crowdsourcing)

How were the data prepared/cleaned? Which procedures were applied?

→ difficult to assess the impact of data preparation if that phase remains opaque



# APIL's Ethics and Big-Data charter (2013)

Purpose: foster better data management

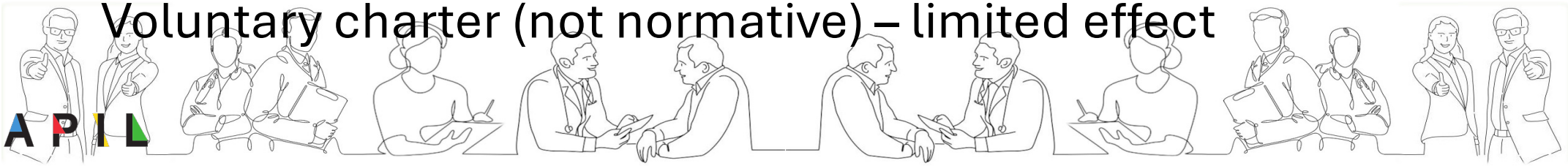
Collaborative work: research, industry, data providers

Three dimensions: before, during and after data processing

Three main sections

- Traceability
- Intellectual property
- Specific Legislation

Voluntary charter (not normative) – limited effect

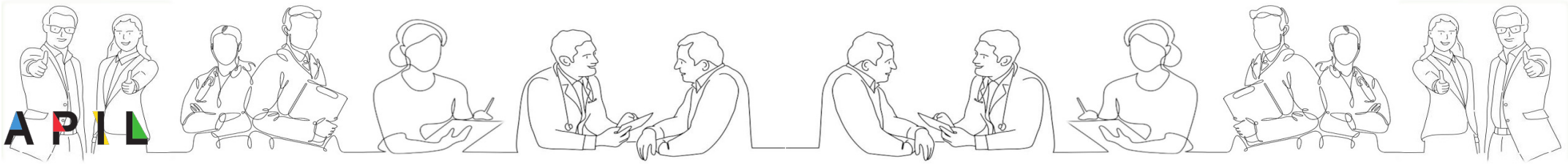


# Current customer requirements: legal

The providers of NLP solutions must ensure that their solutions:

- do not leak any personal/sensitive data
- comply with copyright laws

→ Almost impossible to assess given the opacity of the LLM providers (even the so-called “open-source” ones)



# Current customer technical requirements

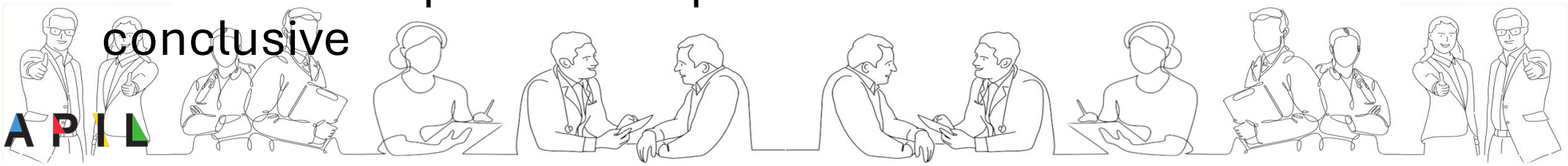
Training and testing

- Data for creating and testing LLMs are built in order to reduce biaises

Technology provider should evaluate the risks

- Overfitting/underfitting
- Hallucination
- Common error patterns, false positives/negatives

→ Too little experience in production LLMs to be conclusive



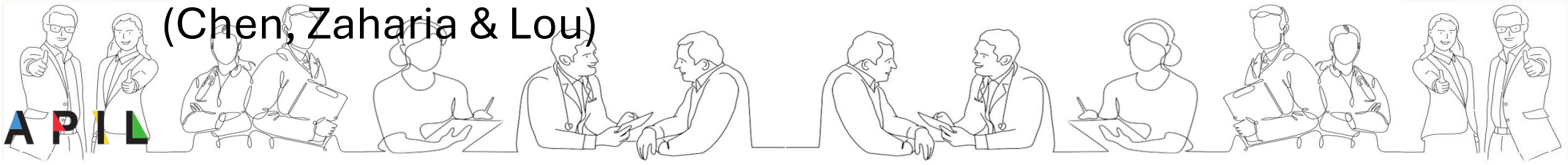
# Current customer requirements: production

Test upgrades

- QA control the impacts of Overfitting/underfitting
- Monitor
- Biases and drifts must be monitored at regular intervals

→Difficult to assess if the product is based on a commercial API with frequent evolutions

e.g. chatGPT is getting worse over time on certain benchmarks (Chen, Zaharia & Lou)



# New regulation: AI Act

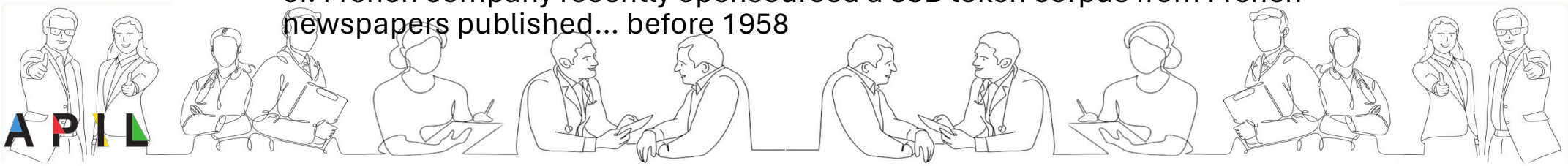
AI Act is putting emphasis on usage risks and on training data copyright, which is a good thing, but actually leaves small actors with two extreme options:

- Use a commercial LLM that embeds copyright problems but also protection

  cf. Microsoft commits to *“defend [its] customers and pay for any adverse judgments if they are sued for copyright infringement”*

- Create their own model which is:
  - Costly,
  - Difficult to do, because of lack of compliant and adequate data

    cf. French company recently opensourced a 85B token corpus from French newspapers published… before 1958



# NLP and LLM: the risks

Industrial

- Cost of assessing compliance to customer expectations
- Mitigate software solution

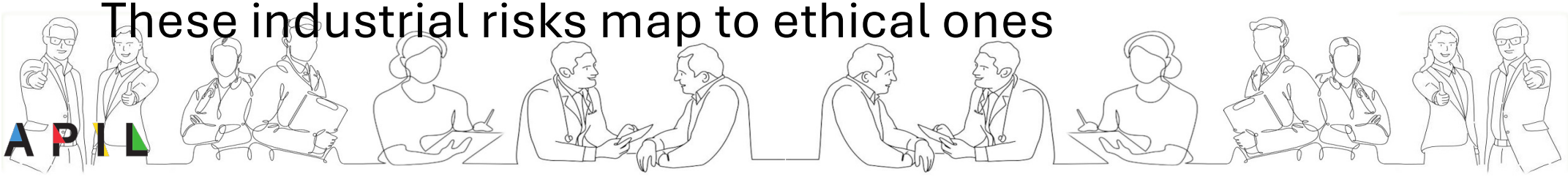
Legal

- Risk of non-compliance (and associated costs)
- Commercial

Reassurance

- Explainability of results

These industrial risks map to ethical ones



# Some ways

Contribute to norms

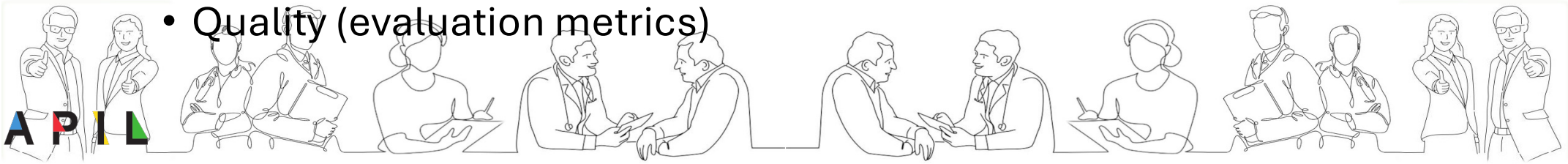
- ISO
- CEN/CENELEC

Describe the data

- Process transparency: sources, transformations
- Annotations (levels and procedures)
- License and copyright

Assess (data and models)

- Compliance
- Quality (evaluation metrics)



# Thank you