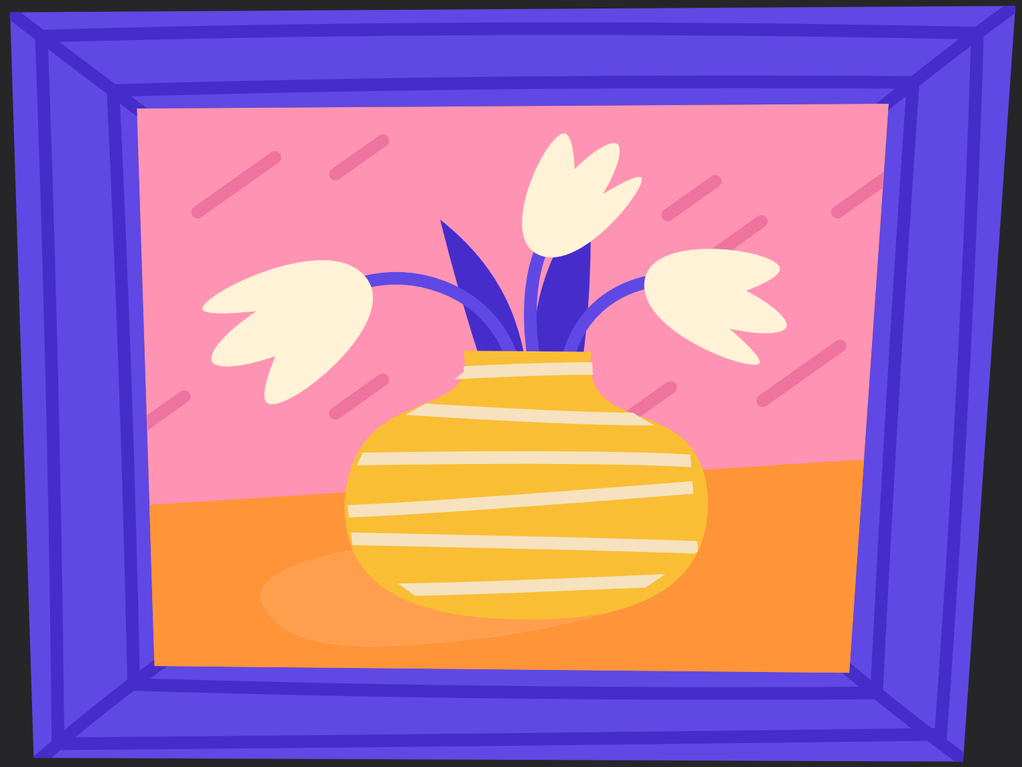
# DIE KUNST DER DATEN

Wie man dem Problem fehlender Daten entgegenwirken kann, wenn man spezialisierte Modelle trainieren will



*Krista Frick & Susanne Waldthaler, Co-Founders wedomagic GmbH*

**René Larch**
Co-Founder & CTO
AI Development
.NET/C#

# UNSER SÜDTIROLER TEAM

Susanne Waldthaler
Co-Founder & CEO
Machine Learning
Python

Krista Frick
Co-Founder & COO
AI implementation
Javascript

# DAS RECHTSPROBLEM

- in Italien sind die gesetzlichen Bestimmungen >10x umfangreicher als Deutschland, Frankreich und UK gemeinsam

- dynamische Gesetzgebung und **vielfältige juristische** Interpretationen

*Quelle: Ufficio studi CGIA, 16.12.2023*

Durchsuchen von Millionen Gesetzen & Urteilen

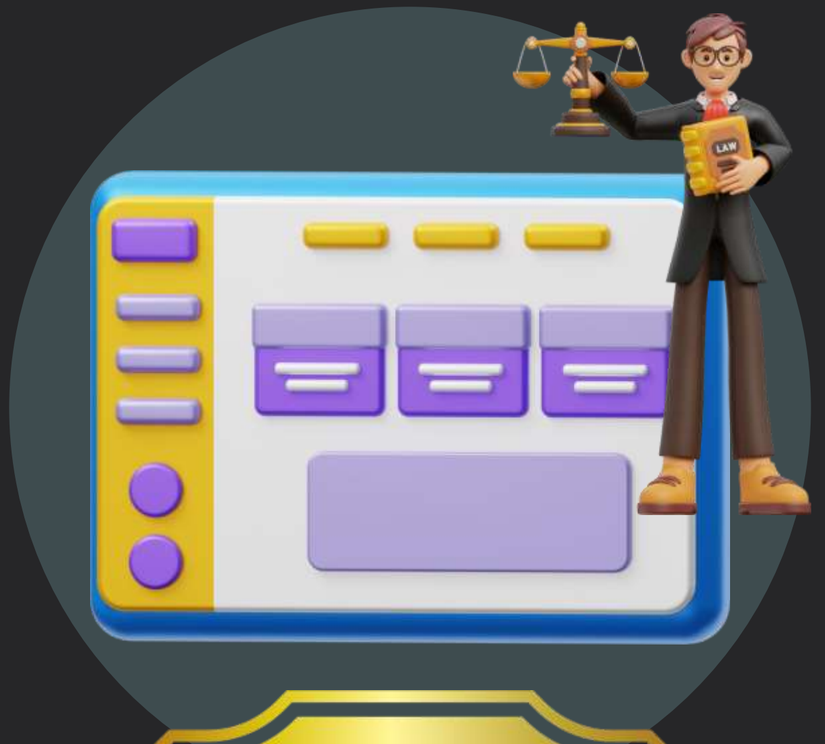

Hoher Zeitaufwand

Altmodische Schlüsselwort-Recherche (Multiple Datenbanken, Bücher)

# UNSERE LÖSUNG

*Mehr Effizienz und Qualität durch eine auf das italienische Recht spezialisierte KI*

- Prüft entsprechende Gesetze und Urteile
- Gibt eine Antwort auf eine juristische Fragestellung im Handumdrehen
- Referenziert strikt verifizierte Quellen (Gesetze & Urteile) und vermeidet “Hallucination”




in Zusammenarbeit mit Anwaltskanzleien entwickelt

# WARUM EIN EIGENES MODELL TRAINIEREN?


- spezialisierte Bereiche brauchen spezialisierte KI-Modelle
- Datenschutz & Privacy
- Zugriff auf Millionen Dokumente (Gesetze, Normen, Rechtssprechung...)

Nach verschiedenen Experimenten und Entwicklungen:

eigenes Frage-Antwort Modell trainieren

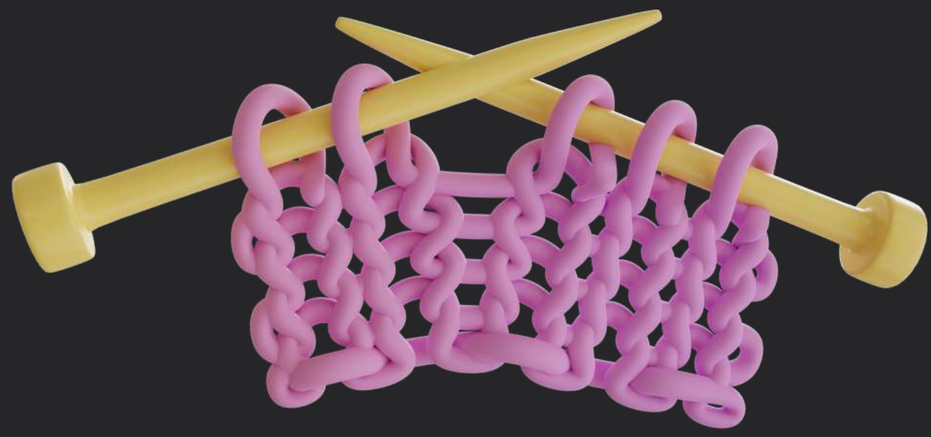
# DIE HERAUSFORDERUNG: TRAININGSDATEN

- Mangel an Daten erschwert das Training von spezialisierten KI-Modellen
- **Rechtliche Beschränkungen** und Datenschutzbestimmungen mit einigen realen Daten
- Hohe Kosten und Aufwand für echte Datengenerierung

# WAS SIND SYNTHETISCHE DATEN?

- Definition: Künstlich generierte Daten, die **echten Daten ähneln/nachahmen**

- Arten von synthetischen Daten:
  - Text (unser Beispiel)
  - Bilder
  - Numerische Daten



# VORTEILE VON SYNTHETISCHEN DATEN

- Überwindung von Datenknappheit
- Vermeidung von Datenschutzproblemen
- Anpassbar und skalierbar
- Kosteneffizient

# PRAKTISCHES BEISPIEL:
# Chatbot für juristische Fragen

1

Endresultat definieren & **Datenverständnis**

2


Modellauswahl je **nach Stärke**

3

Datengenerierung

4

Validierung

# MIXTURE-OF-AGENTS-METHODE

- KI-Agents = KI-Mitarbeiter
- haben unterschiedliche Funktionen und Stärken
- führen zu bestimmten Zeitpunkt - je nach Prozessdefinition - ihre (einzige) Aufgabe aus

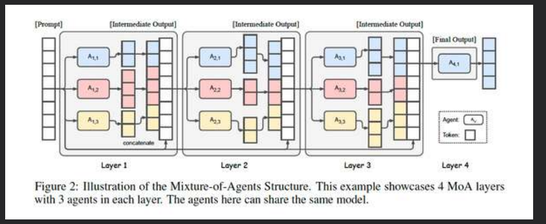


Figure 2: Illustration of the Mixture-of-Agents Structure. This example showcases 4 MoA layers with 3 agents in each layer. The agents here can share the same model.

*Source: Mixture-of-Agents Enahnces Large Language Model Capabilities, Wang et al., Jun 2024, [arXiv:2406.04692](arXiv:2406.04692)*

# CAVEATS UND LEARNINGS

- **Sind synthetische Daten nicht anfällig für** Bias (Verzerrungen)?

  ✔ *Ja! Qualitätskontrolle und Bias-Mitigation-Strategien einsetzen (KI-Agents)! Wo möglich kombinieren mit echten Daten!*

- Wie kontrolliere ich die Richtigkeit der erstellten Daten?

  ✔ *Menschliche Validierung wo möglich (Stichprobenartige Validierung mit Anwälten) + KI-Agents (=KI-Mitarbeiter)*

- Welche Methodik passt zu meiner Aufgabe und meinem Unternehmen?

  ✔ *Guten Plan erstellen und dann...*
  *PROBIEREN GEHT ÜBER STUDIEREN!*

# TAKE HOME MESSAGES

1

GENÜGEND HOCHQUALITATIVE DATEN SIND ESSENZIELL FÜR EFFEKTIVES TRAINING

2

SYNTHETISCHE DATEN KÖNNEN DORT HELFEN WO ECHTE DATEN FEHLEN ODER ZU KOSTSPIELIG SIND

3

ZUSAMMENARBEIT MEHRERER MODELLE IST EIN EFFEKTIVER WEG UM DATEN ZU ERSTELLEN UND **QUALITÄT ZU SICHERN** - MENSCHLICHE VALIDIERUNG!