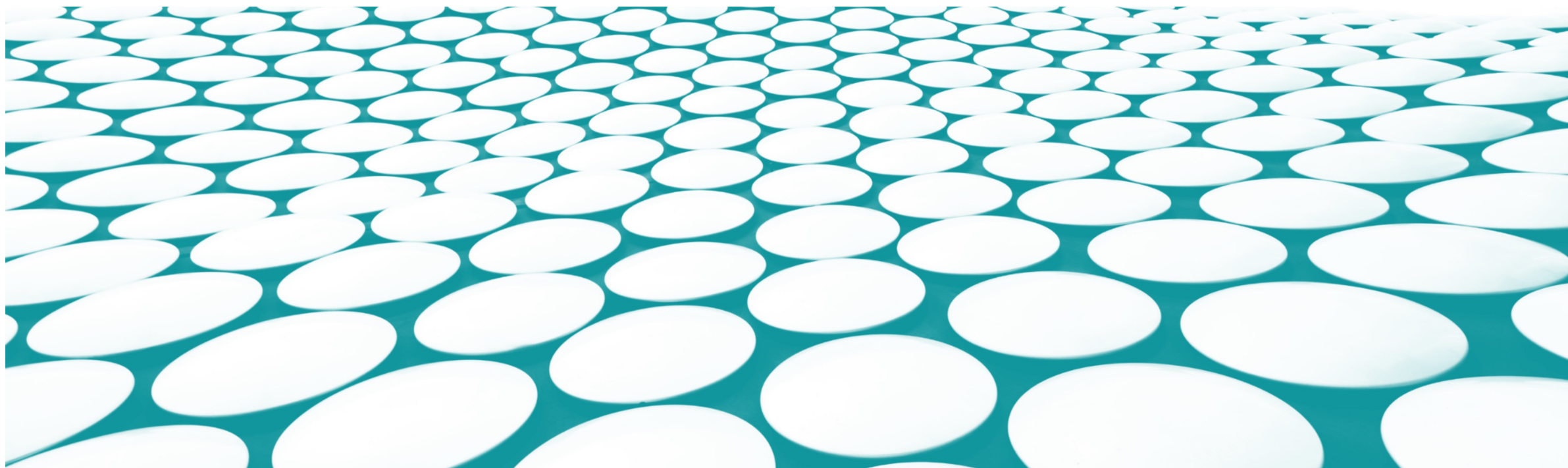




AI LANGUAGE MODEL INNOVATIONS AT NTCA

DR. GÁBOR MAGYAR, CHAIRMAN, NTCA AI WG

14.05.2024





AGENDA

- NTCA **AI** WG – WHO WE ARE?
- A FEW WORDS ON LLMs / SLMs
- EXAMPLE ONGOING NTCA **AI** PROJECTS

NTCA **AI** WG

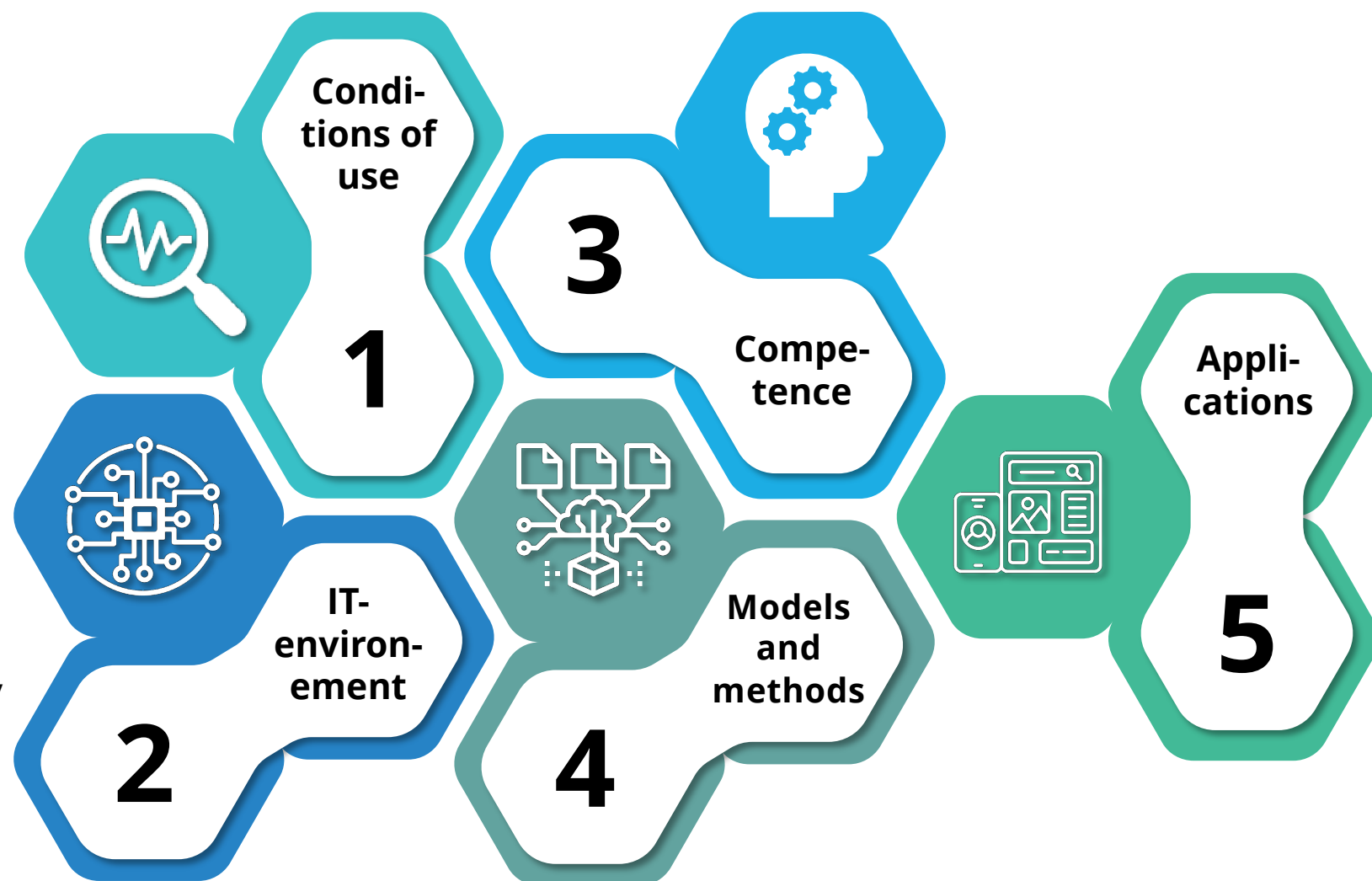
- Established by Government Decree
- The aim of the **AI** WG is
 - to exploit the values existing in the data assets of the NTCA and the State Treasury (as well as in cooperation with other organizations)
 - and thereby put the synergy of scientific methodologies and domain experiences at the service of public good.
- **AI** WG makes proposals, reviews, develops approaches, models and algorithms
 - Eg. concerning to semantic data asset cadastre, teaching/learning models and algorithms, validation methodologies

WHO WE ARE

- Inter-agency working group
- Brings together ministries, gov agencies and academia
- Operates on Tax Agency's IT resources, premises
- Collaborative work in smaller groups by experts from various member institutions and academics
 - Via online collaboration platform
 - Dedicated, protected, physical research environment (connected/merged data analysis)

OUR ROAD WITH **AI**

- **Increase efficiency** in repetitive, routine tasks, so skilled workforce can deal with essential tasks.
- Transition **from declaration-based to event-based operation**: real-time analytics on big data.
- **Serving taxpayers**: adaptive, well-targeted information (deadlines, duties, options).
- **Impact assessments and forecasts**: make the most of data and react faster. Data-rich tax policy analyses.
- **It does not replace domain expertise.**



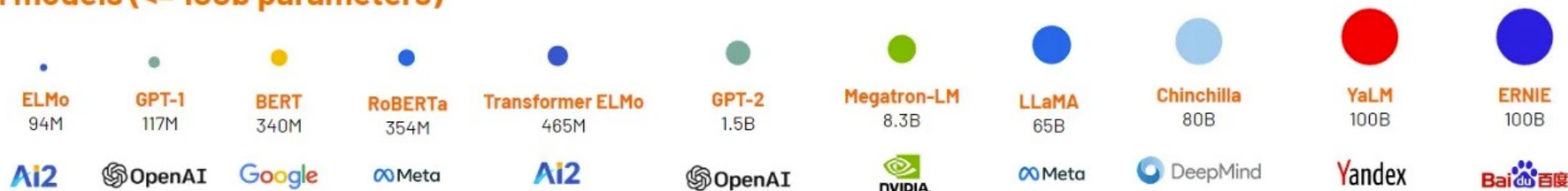
OPPORTUNITIES AND CONSTRAINTS: **AI** AT AN AUTHORITY?

- Precise statutory tasks
- "Hallucination" is not acceptable
- Data security, confidential workflows, liability
- Resource efficiency (human, IT, ...)
- Culture change: social customs, expectations

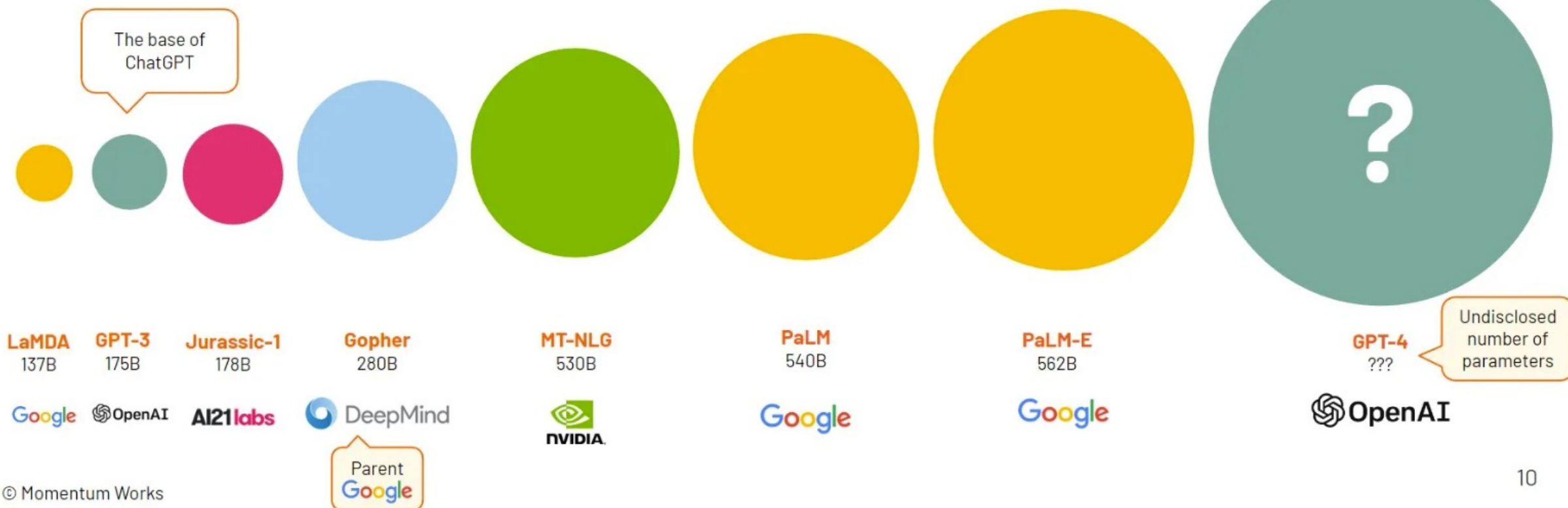
Large Language Models are becoming very large indeed



Small models (<= 100b parameters)



Large models (>100b parameters)



GPT-4o (announced 13/05/2024.)

20 languages were chosen as representative of the new tokenizer's compression across different language families

- English 1.1x fewer tokens (from 27 to 24)
- French 1.1x fewer tokens (from 31 to 28)
- Italian 1.2x fewer tokens (from 34 to 28)
- German 1.2x fewer tokens (from 34 to 29)
- Spanish 1.1x fewer tokens (from 29 to 26)
- Portuguese 1.1x fewer tokens (from 30 to 27)
- Chinese 1.4x fewer tokens (from 34 to 24)
- Japanese 1.4x fewer tokens (from 37 to 26)
- Turkish 1.3x fewer tokens (from 39 to 30)
- Arabic 2.0x fewer tokens (from 53 to 26)
- Persian 1.9x fewer tokens (from 61 to 32)
- Russian 1.7x fewer tokens (from 39 to 23)
- Korean 1.7x fewer tokens (from 45 to 27)
- Vietnamese 1.5x fewer tokens (from 46 to 30)
- Gujarati 4.4x fewer tokens (from 145 to 33)
- Telugu 3.5x fewer tokens (from 159 to 45)
- Tamil 3.3x fewer tokens (from 116 to 35)
- Marathi 2.9x fewer tokens (from 96 to 33)
- Hindi 2.9x fewer tokens (from 90 to 31)
- Urdu 2.5x fewer tokens (from 82 to 33)



SLM AND LLM

Small Language Models (SMLs)

- N*100 million of parameters
- Fine-tuning with a target-specific database

Features:

- Resource efficient
- Task/domain specific

Large Language Models (LLMs)

- Several billions of parameters
- Teaching with instructions and Q&A

Features:

- Zero- and few-shot learning
- A general model for many tasks
- High computing demand
- Use of external data sources (corpus)

[Arena \(battle\)](#)
[Arena \(side-by-side\)](#)
[Direct Chat](#)
[Vision Direct Chat](#)
[Leaderboard](#)
[About Us](#)

LMSYS Chatbot Arena: Benchmarking LLMs in the Wild

[Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)

Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can chat for multiple turns until you identify a winner.
- Votes won't be counted if model identities are revealed during the conversation.

Chatbot Arena [Leaderboard](#)

- We've collected 1,000,000+ human votes to compute an Elo leaderboard for 90+ LLMs. Find out who is the 🏆 LLM Champion!

Chat now!

Expand to see the descriptions of 39 models

[Claude](#): Claude by Anthropic

[Reka Core](#): Frontier Multimodal Language Model by Reka

[Qwen Max](#): The Frontier Qwen Model by Alibaba

[Qwen 1.5](#): The First 100B+ Model of the Qwen1.5 Series

[Llama 3](#): Open foundation and chat models by Meta

[Gemini](#): Gemini by Google

[Snowflake Arctic Instruct](#): Dense-MoE transformer by Snowflake AI

[Phi-3](#): A capable and cost-effective small language models (SLMs) by Microsoft

[Mixtral of experts](#): A Mixture-of-Experts model by Mistral AI

[GPT-4-Turbo](#): GPT-4-Turbo by OpenAI

[GPT-3.5](#): GPT-3.5-Turbo by OpenAI

[Reka Flash](#): Multimodal model by Reka

[Command-R-Plus](#): Command-R Plus by Cohere

[Command-R](#): Command-R by Cohere

[Gemma](#): Gemma by Google

[Zephyr 141B-A35B](#): ORPO fine-tuned of Mixtral-8x22B-v0.1

[DBRX Instruct](#): DBRX by Databricks Mosaic AI

[Llama 2](#): Open foundation and fine-tuned chat models by Meta



✂ LMSYS Chatbot Arena: Benchmarking LLMs in the Wild

- [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)



Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can chat for multiple turns until you identify a winner.
- Votes won't be counted if model identities are revealed during the conversation.



Chatbot Arena [Leaderboard](#)

- We've collected 1,000,000+ human votes to compute an Elo leaderboard for 90+ LLMs. Find out who is the 🏆 LLM Champion!



Chat now!

🔍 Expand to see the descriptions of 39 models

Model A

How reliable an LLM-based chat agent can be, in case it is intended to be used in a strictly legal rule-based customer relationship job?

The reliability of an LLM-based chat agent in a strictly legal rule-based customer relationship job would depend on several factors:

1. **Training data:** The quality and relevance of the data used to train the LLM is crucial. It should include a comprehensive set of legal rules, regulations, and case studies specific to the domain in which the chat agent will operate.
2. **Fine-tuning:** The LLM should be fine-tuned on the specific legal domain to ensure it can provide accurate and relevant responses to customer queries. This may involve training on a smaller, domain-specific dataset.
3. **Continuous learning and updates:** As legal rules and regulations change over time, the LLM should be regularly updated with the latest information to maintain its reliability.

Model B

How reliable an LLM-based chat agent can be, in case it is intended to be used in a strictly legal rule-based customer relationship job?

Large Language Models (LLMs) like myself can be quite versatile and powerful tools for a variety of applications, including customer service roles that involve legal or rule-based interactions. However, their reliability in a strictly legal setting depends on several factors:

1. **Training Data:** The quality and specificity of the training data are crucial. If the model has been trained on a large corpus that includes extensive legal texts, regulations, and precedent cases, it will be better equipped to handle legal inquiries.
2. **Fine-Tuning:** Additional fine-tuning on domain-specific legal data can significantly improve the model's accuracy and relevance to the legal field. This process tailors the model to understand and respond according to legal norms and terminology.
3. **Contextual Understanding:** While LLMs excel at understanding context, they may sometimes struggle with the nuances of complex legal issues or very specific, rare cases.



✂ LMSYS Chatbot Arena: Benchmarking LLMs in the Wild

◦ [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)

📖 Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can chat for multiple turns until you identify a winner.
- Votes won't be counted if model identities are revealed during the conversation.

🏆 Chatbot Arena [Leaderboard](#)

- We've collected 1,000,000+ human votes to compute an Elo leaderboard for 90+ LLMs. Find out who is the 🏆 LLM Champion!

👉 Chat now!

🔍 Expand to see the descriptions of 39 models

Model A

How would you create a domain-specific chat model that has the power of an LLM, but the specific text corpus can stay in a confidential, closed frame at the operator?

To create a domain-specific chat model that has the power of an LLM while keeping the specific text corpus confidential and closed at the operator level, you can follow these steps:

1. Choose a pre-trained LLM: Select a powerful pre-trained language model, such as GPT-3, BERT, or T5, which has been trained on a large, diverse corpus of text data. These models have a strong understanding of language and can generate coherent responses.
2. Prepare your domain-specific corpus: Collect and preprocess the text data specific to your domain. This could include legal documents, case studies, customer interactions, and other relevant materials. Ensure that the data is clean, well-structured, and free of sensitive information.
3. Fine-tune the LLM: Use transfer learning techniques to fine-tune the pre-trained LLM on your domain-specific corpus. This process involves training the model on your data while

Model B

How would you create a domain-specific chat model that has the power of an LLM, but the specific text corpus can stay in a confidential, closed frame at the operator?

Creating a domain-specific chat model with the capabilities of a Large Language Model while keeping the training data confidential and within a closed system involves a few key steps. Here's a high-level overview of the process:

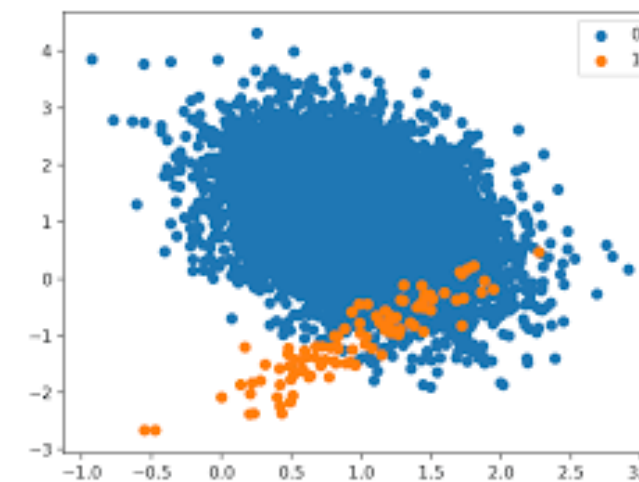
1. **Data Collection & Preparation:** Gather your domain-specific legal documents, previous customer interactions, regulations, and any other relevant text corpus. Ensure this data is properly anonymized and complies with data protection regulations. Since confidentiality is a concern, make sure all data handling adheres to internal security protocols.
2. **On-Premise or Private Cloud Setup:** To keep the data within a closed frame, you'll need to set up an on-premise server or use a private cloud infrastructure where you have full control over access and security. This setup ensures that your data does not leave your



EXAMPLE ONGOING PROJECTS AT NTCA **AI** WG

NOWCASTING BASED ON RETAIL DATA

- **GOAL:**
time-series analysis of real-time retail data (impact assessment, forecasts, decision support)
- Analysis of online cash register data using machine learning methods
- Cleaning and classification of product names are required:
normalization and hiding company-specific data
- Building in-house experience with AI/deep learning-based methods



NOWCASTING BASED ON RETAIL DATA



Targets

Automated processing of real-time product data (time series analysis, economic impact assessment, inflation now-casting)



Data

Product names recorded in online cash register .

1.5 million unique product names



Partners

NTCA,
National Bank of Hungary

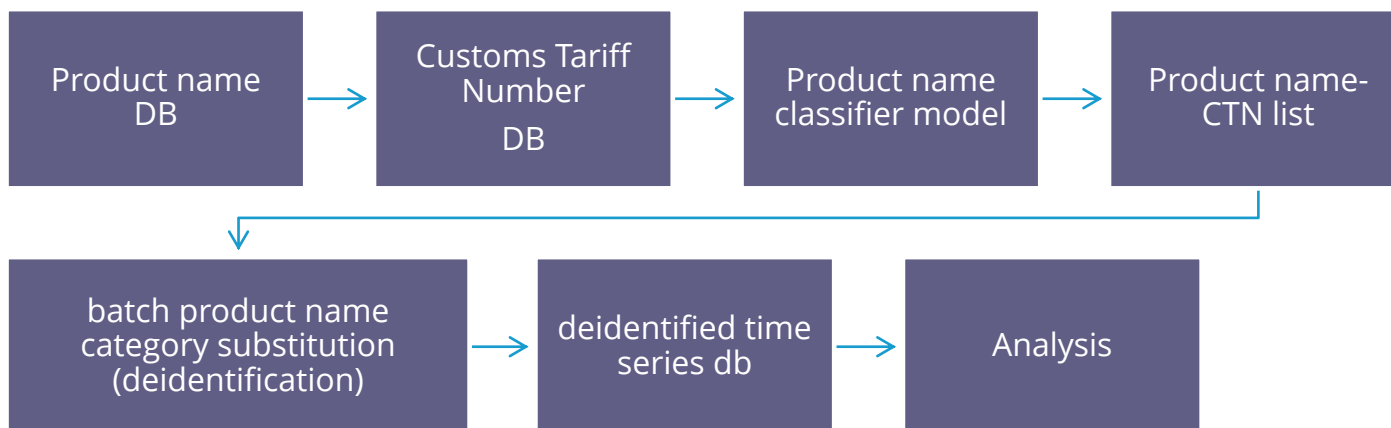
Central Statistical Office

University researchers,

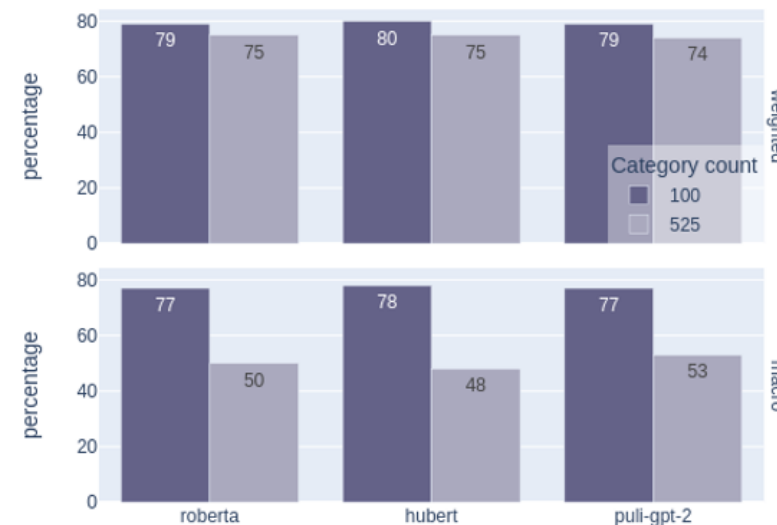


Methods

Antecedent: tf-idf
Today: ttransformer-based pre-trained language model (to vectorize texts). Classification: different procedures (decision tree, random forest, XGBoost, MLP, etc.). Train the upper layers of a transformer-based pre-trained language in model classifications.



Finetuned model accuracies

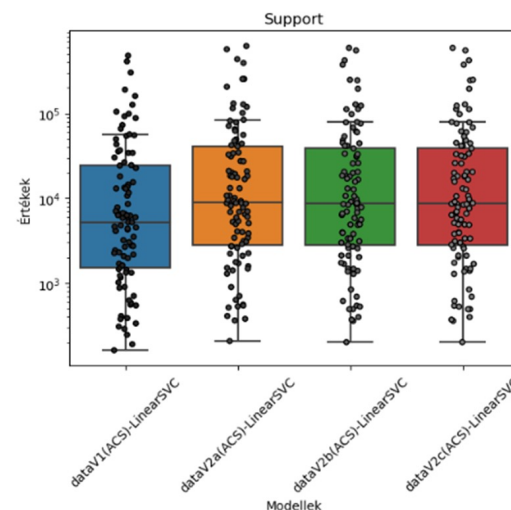
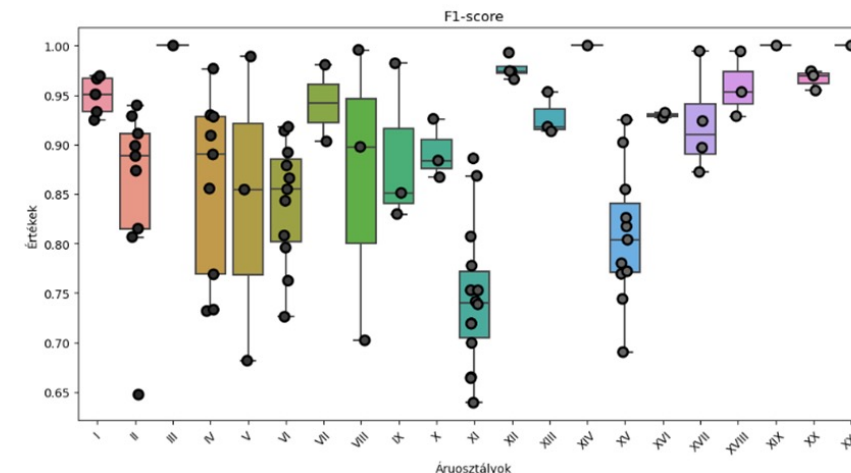


ONLINE INVOICE ITEM CATEGORIZATION

GOAL:

recognition of product or service content in online invoices,
and matching to tariff codes.

- Data extraction
- Discover and correct
 - false and outdated tariff classification
 - false service codes

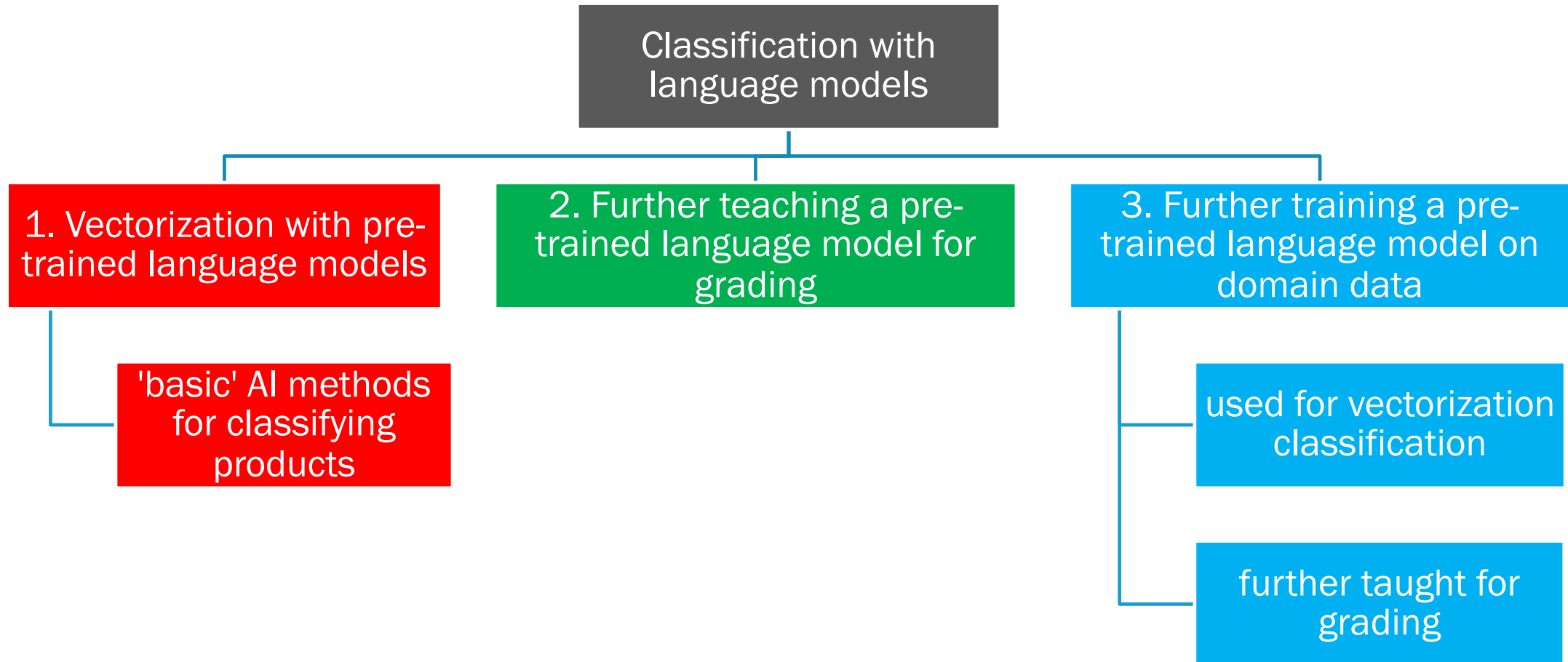


Multi data-set input:

- Online invoice
- Online cash register
- Customs export/import data
- Public Road Trade Control data
- Taric db, code lists



THE ROLE OF **AI**: ITEMS CATEGORIZATION USING DEEP LEARNING-BASED METHODS



CHECKING FOR "UNUSUALNESS" OF INVOICE DATA

The goods and services described on the invoices can be recognized „unusual” in some way (related to the taxpayers involved).

In service to taxpayers: drawing attention to the unusual nature of the invoice data
risk analysis: faster response, emerging new risk indicators

„Unusual”

- in values
- by filed business activity
- in time-patterns
- by unusualness of the good/service
- e.t.c.

- DESCRIPTIVE STATISTICAL METHODS
- Statistical analysis of the taxpayer's previous invoices („habits”)
- Expert methods for clustering
- **AI-BASED METHODS**
- Supervised classification
- Unsupervised classification
- Account "expected value" estimation (supervised regression)

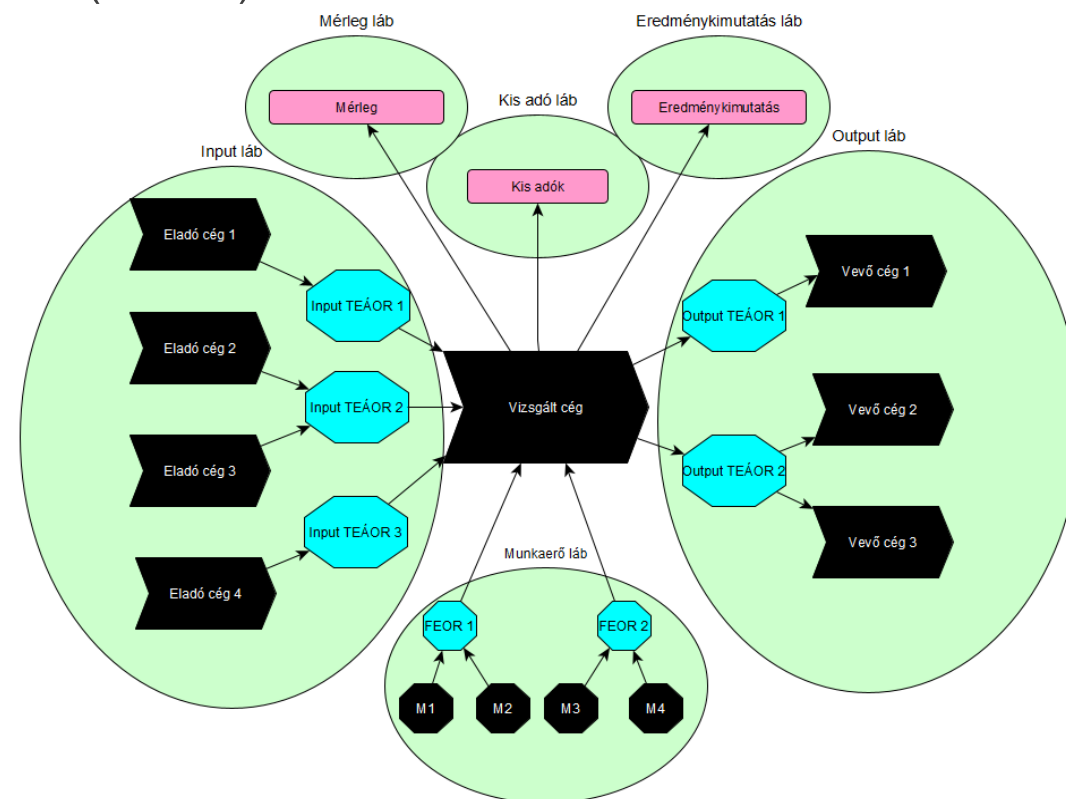
MAIN ACTIVITY (TEÁOR COMPLIANCE)

GOAL:

Verification of the adequacy of the declared main activity and economic activity data. *Recommendation app* of TEÁOR (NACE)

Based on

- Accounts received
- Recipients of issued invoices
- Workforce admin data
- Sectoral taxes characterising activities



ADMINISTRATIVE COST ESTIMATION

GOAL: a time-series estimation (at national economy level) of the trends of administrative costs in business activities (primarily related to tax activities).

Development of a methodology,
based solely on administrative data.
Time series estimation at national
level

Based on accounting,
payroll and consulting
account data,
as well as contribution
declarations for specialised
employees

**THXs for your
attention!**

