Generative AI - opportunities, risks and challenges

Andrejs Vasiljevs
Co-Founder and Board Member, Tilde
Senior researcher, University of Latvia
Interest in Generative AI by online searches

Source: Google Trends
Interest in Large Language Models by online searches

Source: Google Trends

Nov 30, 2022
Release of ChatGPT

Source: Google Trends
Rapid rise of Large Language Models

Chronological display of LLM releases

The trends in the number of LLM models introduced over the years

Hype Cycle for Artificial Intelligence, 2023

Gartner
Variety of tasks performed by Generative AI

- Text Creation
- Music Composition
- Image Generation
- Code Generation
- Voice Generation
- Deepfakes
- Animation
- Drug Discovery
- Simulation Environments
- Marketing Materials
- Personalised Experiences
- Tailored Education
- Data Augmentation
- Video Dubbing

Personalised Experiences

Tailored Education

Data Augmentation

Video Dubbing

Deepfakes

Animation

Drug Discovery

Simulation Environments

Marketing Materials
Variety of tasks performed by Large Language Models

- Text Generation
- Conversation
- Translation
- Summarization
- Question Answering
- Sentiment Analysis
- Named Entity Recognition
- Text Classification
- Part-of-Speech Tagging
- Grammar and Spell Checking
- Writing Assistance
- Code Generation and Debugging
- Content Recommendation
- Semantic Search
- Tutoring and Education
- Text-to-Speech and Speech-to-Text
- Data Extraction and Processing
- Creative Writing and Ideation
- Interactive Storytelling
- Dialogue Frameworks for Virtual Assistants
Over 100 electronic government services in Latvia are utilizing advanced chatbots, which are enhanced by large language models, for their customer service functions.
LLMs as part of National Language Technology Infrastructure
Hype Cycle for Artificial Intelligence, 2023

- Generative AI Foundation models

Gartner
Language disparity in Large Language Models

Credits: Marta Villegas, BSC-CNS
Creating Large Language Models in EU member states

• **Language Technology Initiative project in Latvia** - creating synergy between higher education, science and industry to promote innovation in language technology:
  
  • Preparing specialists in language technology
  
  • Creating **Large Language Models for Latvian** and other advanced language technologies

• Large Language Model development projects in Germany, Spain, France, Finland, Norway, Netherlands, Portugal, Estonia and other EU member states
Extensive approach in training data for Large Language Models

Credits: Marta Villegas, BSC-CNS
Except for English and, to a certain extent, German, Spanish and French, all European languages are massively under-resourced in terms of data, tools and technologies!

Credits: Georg Rehm, DFKI / Language Data Space
Previous EU Projects and Initiatives in Language Data

• **META-NET** (FP7, 2010-2013)
  • META-SHARE

• **ELRC** (CEF, 2014-2023)
  • ELRC-SHARE

• **ELG** (H2020, 2019-2022)
  • ELG Cloud Platform

• **ELE** (PP/PA, 2021-2023)

*Credits: Georg Rehm, DFKI / Language Data Space*
Objective
Develop and deploy a European platform and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data

Salient features
Governance framework, technical architecture and infrastructure, openness, promotion

Stakeholders
Industry, research, public administration, cultural associations, NGOs and citizens

Sustainability
Towards long-term sustainability: ongoing work on supporting EU Member States in creating ALT-EDIC (Alliance for Language Technology European Data Infrastructure Consortium)
LLMs for Europe made in Europe – The Big Picture

**LLM Platform**
- User-friendly platform for pre-training, fine-tuning and deployment of LLMs
- Can be realised on top of ELG (requires additional development work & HPC access)
- Huge gap that needs to be addressed soon

**NLP/LT Platform**
- General NLP and Language Technology Platform
- Host and provide general, i.e., non-LLM-based, NLP and LT services
- Addressed by European Language Grid

**Federated Learning**
- Federated learning of LLMs by making use of the data sets provided in a decentralised way by LDS
- Huge gap that needs to be addressed soon

**Data Curation**
- High-performance tools for language data curation and annotation
- Filtering, quality and bias assessment, identification and analysis of various text and document properties (language, genre, structure etc.)
- Huge gap that needs to be addressed soon

**Computing Infrastructure**
- High-Performance Compute infrastructure for training LLMs
- Addressed by EuroHPC Joint Undertaking
- Simple yet crucial issue: storage capacity, esp. for multimodal data

**Language Data**
- Secure and trusted sharing of language data covering all European languages
- Decentralised, i.e., federated
- Addressed by Language Data Space

Credits: Georg Rehm, DFKI / Language Data Space
Strategies for Advancing Generative AI in Europe

Europe should invest heavily in research, education, and infrastructure while implementing balanced regulations to foster innovation and public trust in Generative AI:

• Ensuring that Large Language Models fully represent linguistic and cultural diversity of Europe.

• Fostering research towards smaller and greener models that are as efficient as (or better than) huge resource-consuming models like GPT-4.

• Making R&D and innovation in LLMs and Generative AI as open and accessible to large spectrum of European players as possible, including innovative startups and SMEs:
  • Availability of globally competitive HPC infrastructure
  • Availability of voluminous and diverse data
  • Broad, powerful and easy to use enabling services
Thank you!

andrejs@tilde.com