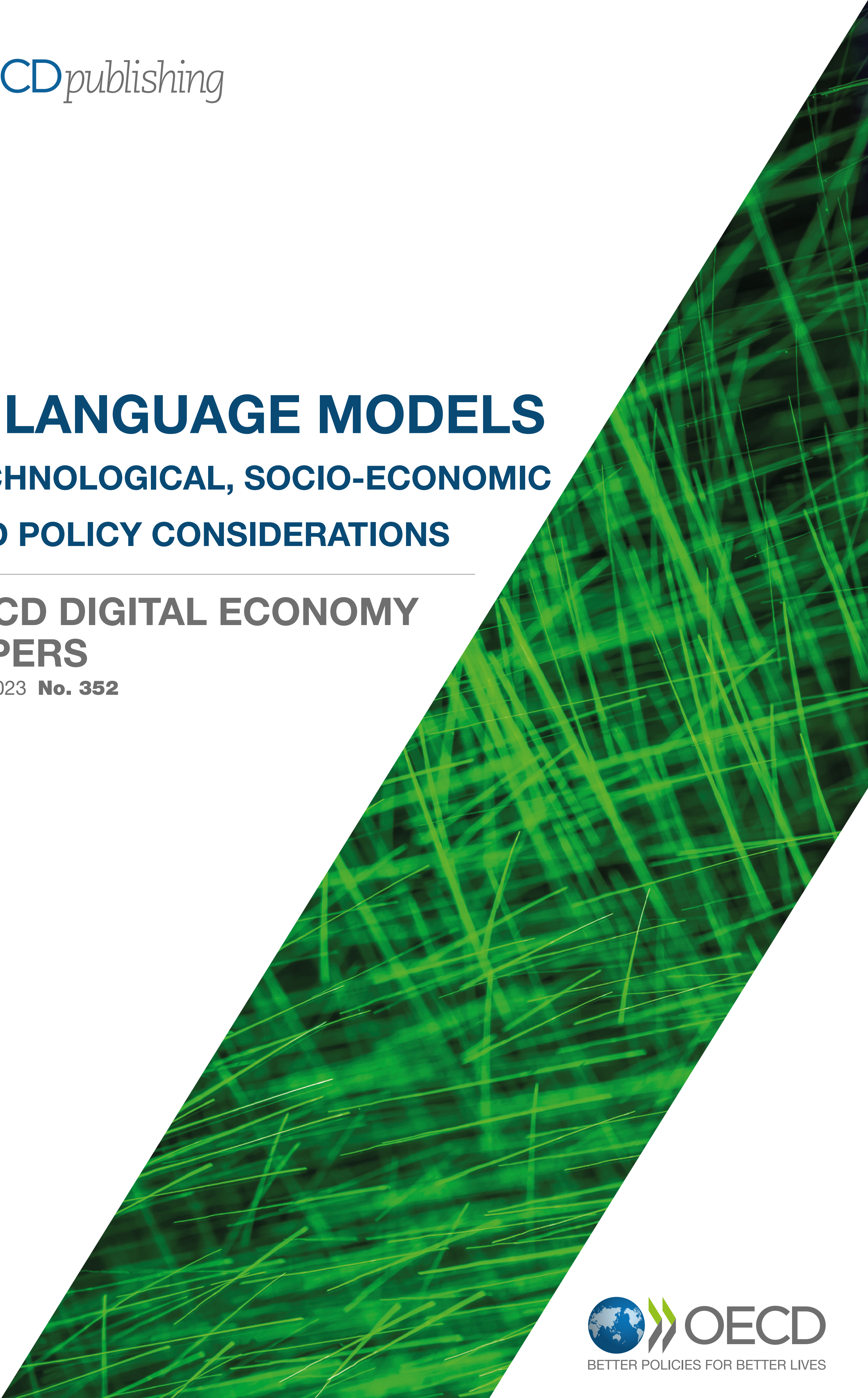
OECDpublishing

AI LANGUAGE MODELS

TECHNOLOGICAL, SOCIO-ECONOMIC AND POLICY CONSIDERATIONS

OECD DIGITAL ECONOMY PAPERS

April 2023 **No. 352**



OECD
BETTER POLICIES FOR BETTER LIVES

This paper was prepared for publication by the OECD Secretariat in consultation with the delegates of the Working Party on Artificial Intelligence Governance (AIGO). The paper was approved and declassified by written procedure by the Committee on Digital Economy Policy (CDEP) on 10/03/2023.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP/AIGO(2022)1/FINAL

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city, or area. The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem, and Israeli settlements in the West Bank under the terms of international law.

Abstract

AI language models are a key component of natural language processing (NLP), a field of artificial intelligence (AI) focused on enabling computers to understand and generate human language. Language models and other NLP approaches involve developing algorithms and models that can process, analyse, and generate natural language text or speech trained on vast amounts of data using techniques ranging from rule-based approaches to statistical models and deep learning. The application of language models is diverse and includes text completion, text-to-speech conversion, language translation, chatbots, virtual assistants, and speech recognition. This report offers an overview of the AI language model and NLP landscape with current and emerging policy responses from around the world. It also explores the basic building blocks of language models from a technical perspective using the OECD Framework for the Classification of AI Systems. Finally, the report presents policy considerations associated with AI language models through the lens of the OECD AI Principles.

Abrégé

Les modèles de langage d’intelligence artificielle (IA) sont un élément clé du traitement du langage naturel (NLP), un domaine de l’IA qui vise à permettre aux ordinateurs de comprendre et de générer le langage humain. Le NLP implique le développement d’algorithmes et de modèles capables de traiter, d’analyser et de générer des textes ou des discours en langage naturel. Ils sont formés sur de grandes quantités de données, en utilisant des techniques allant des approches basées sur des règles aux modèles statistiques et à l'apprentissage en profondeur. Les applications des modèles de langage sont diverses et comprennent notamment l’auto-complétion de texte, la conversion texte-parole, la traduction linguistique, les *chatbots*, les assistants virtuels et la reconnaissance vocale. Ce rapport offre une vue d’ensemble du paysage des modèles de langage IA et du NLP avec les réponses politiques actuelles et émergentes du monde entier. Il explore également les éléments de base du NLP d’un point de vue technique en utilisant le cadre de l’OCDE pour la classification des systèmes d’IA. Enfin, le rapport présente des considérations politiques associées aux modèles de langage IA à travers le prisme des Principes sur l’IA de l’OCDE.

Acknowledgements

The paper “AI language models: Technological, socio-economic and policy considerations” was prepared under the aegis of the OECD Working Party on Artificial Intelligence Governance (AIGO) and the OECD Committee for Digital Economy Policy (CDEP). AIGO delegates and national contact points for the database of national AI policies of the OECD.AI Policy Observatory contributed significantly.

This report was a collective effort co-ordinated by the AI Unit in the OECD Digital Economy Policy division. Research, interviews, stock-taking and drafting were conducted by Andras Hlacs, Sharon Ho and Laura Galindo. Analysis and writing were conducted by Karine Perset and Luis Aranda.

The team gratefully acknowledges the contributions made by individuals and institutions that participated in expert interviews: Sebastian Hallensleben (CEN-CENELEC JTC 21); Amir Banifatemi (AI Commons); Alexandra Belias, Conor Griffin, and Laura Weidinger (DeepMind); Jan Hajic, Katrin Marheinecke, Georg Rehm, and German Rigau (European Language Grid); Mathilde Bras, Margaret Mitchell, Carlos Muñoz Ferrandis, Giada Pistilli, Anna Rogers, Yacine Yernite, and Thomas Wolf (Hugging Face); Antonella Zarra (META); Alexander Huggins (Nuance); Ferenc Kása and László Boa (Digitalis Jolet Nonprofit); and Mark A. Finlayson (Florida International University).

In addition, the quality of the report benefited significantly from the engagement and written contributions of Stuart Russell (University of California, Berkeley), Carlos Muñoz Ferrandis (Hugging Face), Nobuhisa Nishigata and Takayuki Honda (Ministry of Internal Affairs and Communications, Japan), Miguel Valle Del Olmo (Spanish Permanent Representation to the EU), Yasmin Azhari (Permanent Delegation of Israel to the OECD), Sarah Box (Belgotex New Zealand), Christine Hafskjold (Norwegian Ministry of Local Government and Regional Development), Suhyeon Kang (Ministry of Science and ICT, Korea), Katya Klinova (Partnership on AI), Mohammed Motiwala (United States Department of State), Luis Ricardo Sánchez Hernández (INAI, the National Data Protection Authority in Mexico), Arnis Skraučs (Ministry of Environmental Protection and Regional Development of the Republic of Latvia), Prateek Sibal (UNESCO), and Gustavo Ziemath (Ministry of Foreign Affairs of Brazil).

Finally, the team acknowledges the helpful input and reviews by OECD team members: Celine Caira, Jamie Berryhill, Alistair Nolan, Giuseppe Bianco and Andras Molnar. The team also greatly thanks John Tarver for his help editing this report and Orsolya Dobe, Andreia Furtado, Angela Gosmann, and Sebastian Ordelheide for editorial and publishing support.

Table of contents

FIGURES

TABLES

BOXES

Acronyms and abbreviations

AI	Artificial intelligence
ASR	Automatic speech recognition
EU	European Union
IP	Intellectual property
IGO	Intergovernmental organisation
LM	Language model
LLMs	Large language models
LT	Language technology
NLP	Natural language processing
NLPL	Nordic Language Processing Laboratory
OECD	Organisation for Economic Co-operation and Development
RNN	Recurrent neural network
R&D	Research and development
S2T	Speech-to-text
SLMs	Smaller language models
T2S	Text-to-speech
UNESCO	United Nations Educational, Scientific and Cultural Organisation

Executive summary

AI language models are part of an increasingly critical subset of AI technologies known as natural language processing (NLP). NLP uses language as an input, produces language as an output, or both. Although deployment is at a relatively early stage, language models are widely viewed as transformative, as evidenced by rapid growth in investment and widespread adoption of applications such as the conversational agent ChatGPT.

AI language models promise to unlock significant opportunities to benefit people by conducting tasks in human natural language at scale.

AI language models are being deployed across sectors, such as public administration, healthcare, banking, and education, boosting productivity and decreasing costs. They enable language recognition, interaction support and personalisation. They also enable interactive dialogue systems and personal virtual assistants. AI language models can help safeguard minority or endangered languages by allowing them to be heard, taught, and translated.

Policy makers want an enabling policy environment while mitigating the risks of AI language models.

As individuals and organisations integrate language models into their functioning and services, questions are being raised about how policy makers can ensure that these transformative models are beneficial for people, inclusive and safe. The 2019 OECD AI Principles state that “AI systems should be robust, secure and safe throughout their entire life cycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.” Yet increasingly powerful AI language models raise significant policy challenges related to their trustworthy deployment and use.

AI language model need quality control and standards to address issues of opacity, explainability, accountability and control.

Many AI language models use neural networks that are opaque and complex. The lack of understanding of their internal principles of operation and how they reach specific outputs, even by those who develop them, leads to unpredictability and inability to constrain behaviour. Policy makers must encourage all actors, notably researchers, to develop rigorous quality control methodologies and standards for systems to meet, appropriate to the application context.

The complexity of language models also means that it can be difficult and costly to understand which parties and what data are involved in their development, training, and deployment to enable accountability by those best able to mitigate specific risks. In addition, many language technologies in use today leverage pre-existing models into which users have little visibility. Human overreliance on the outputs of language models is another risk to accountability.

AI language models pose risks to human rights, privacy, fairness, robustness, security, and safety.

AI language models are a form of “generative AI”. Generative AI models create new content in response to prompts, based on their training data. The training data itself can include biases, confidential information on individuals and information associated with existing intellectual property claims and rules. Language models can then discriminate, leak or infer confidential or rights-infringing information.

AI language models can help actors manipulate opinions at scale and automate mis- and disinformation in a way that can threaten democratic values, which is particularly challenging.

AI language models can facilitate and amplify the production and distribution of fake news and other forms of manipulated language-based content that may be impossible to distinguish from factual information, raising risks to democracy, social cohesion, and public trust in institutions. AI “hallucinations” also occur, where models generate incorrect outputs but articulate them convincingly. The combination of AI language models and mis- and disinformation can lead to deception at a wide scale that traditional approaches like fact checking, detection tools, and media literacy education cannot readily address.

Continued dialogue and research can help to mitigate the risks of complex language models.

Further research is essential to understand these complex models and find risk mitigation solutions. In addition, looking to the future, guardrails may be called for to control some forms of powerful language models that can directly affect the real world. One question with significant societal implications remains: should powerful language models be able to take actions directly, such as sending emails, making purchases, and posting on social media, as opposed to their current use as passive question-answering systems?

Language and computing resources as well as language models are key.

The limited availability of digitally readable text to train models remains an important issue for many languages. The most advanced language-specific models use the languages for which significant digital content is available, such as English, Chinese, French, or Spanish. Policy makers in countries with minority languages are promoting the development of digital language repositories, plans, and research. Multilingual language models can foster inclusion and benefit a broader range of people. Access to computing hardware is also crucial but needs more R&D to reduce financial and environmental costs in favour of more efficient mechanisms to train and query large language models.

To ensure the benefits of language technologies are widely shared, actors will have to prepare for economic transitions and equip people with skills to develop and use AI language models.

Language models have the potential to automate tasks in many job categories. AI language models such as OpenAI’s Generative Pre-trained Transformer 4 (GPT-4) are increasingly used to help perform tasks previously conducted by people, including high-skill tasks such as writing software code, drafting reports and even creating content. GPT-4 exhibits human-level performance across several standardised tests. Policy makers will have to experiment with new social models and re-evaluate education needs in an era of ubiquitous AI language models.

International, interdisciplinary, and multi-stakeholder cooperation for trustworthy AI language models is required to address harmful uses and impacts.

Stakeholders, including policy makers, are beginning to explore related societal impacts and risks. Collaboration is taking the form of sharing best practices and lessons learned in regional and international fora and developing joint initiatives in multilingual language data and models. Yet, work remains to develop viable policy and technical solutions that can effectively mitigate risks from language models and other types of generative AI while fostering their beneficial development and adoption. All actors in the AI ecosystem have key roles to play.

Résumé

Les modèles de langage de l’IA font partie d'un sous-ensemble de plus en plus important de technologies d'IA, connu sous le nom de traitement automatique du langage naturel (NLP). Le NLP utilise le langage comme source d'information et produit du langage comme résultat, ou les deux à la fois. Bien que leurs déploiements n'en soient qu'à leurs débuts, les modèles de langage sont largement considérés comme étant transformateurs, comme en témoignent la croissance rapide des investissements et l'adoption généralisée d'applications telles que l'agent conversationnel ChatGPT.

Les modèles de langage de l'IA offrent des possibilités considérables d'améliorer la vie des gens en effectuant des tâches en langage naturel humain à grande échelle.

Les modèles de langage d'IA sont déployés dans des secteurs tels que l'administration publique, les services de santé, les services financiers et l'éducation, pour accroître la productivité et réduire les coûts. Ces modèles permettent la reconnaissance de la langue, l'aide à l'interaction et la personnalisation. Ils permettent également de mettre en place des systèmes de dialogue interactif et des assistants virtuels personnels. Les modèles de langage d’IA peuvent contribuer à la sauvegarde des langues minoritaires ou menacées en leur permettant d'être entendues, enseignées et traduites.

Les décideurs politiques veulent un environnement politique favorable tout en atténuant les risques liés aux modèles de langage d'IA.

Alors que les individus et les organisations intègrent les modèles de langage dans leur fonctionnement, des questions se posent sur la manière dont les responsables politiques peuvent s'assurer que ces modèles transformationnels sont bénéfiques pour les personnes, inclusifs et sûrs. Les Principes sur l'IA de l'OCDE stipulent que « Les systèmes d’IA devraient être robustes, sûrs et sécurisés tout au long de leur cycle de vie, de sorte que, dans des conditions d’utilisation normales ou prévisibles, ou en cas d’utilisation abusive ou de conditions défavorables, ils soient à même de fonctionner convenablement, et ne fassent pas peser un risque de sécurité démesuré. » Pourtant, des modèles de langage d'IA de plus en plus puissants soulèvent d'importants enjeux de politiques publiques liés à leur déploiement et à leur utilisation en toute confiance.

Les défis posés par les modèles de langage d'IA nécessitent un contrôle de la qualité et des normes en matière de transparence, d'explicabilité, de responsabilité et de contrôle.

De nombreux modèles de langage d'IA utilisent des réseaux de neurones opaques et complexes. Le manque de compréhension de leurs principes internes de fonctionnement et de la manière dont ils obtiennent des résultats spécifiques, même par ceux qui les ont développés, conduit à l'imprévisibilité et à l'incapacité de contraindre les comportements. Les décideurs politiques doivent encourager tous les acteurs, notamment les chercheurs, à développer des méthodologies et des normes rigoureuses de contrôle de la qualité adaptées au contexte d'application des systèmes.

La complexité des modèles linguistiques signifie également qu’il peut être difficile et coûteux de comprendre quelles parties et quelles données sont impliquées dans leur développement, leur formation et leur déploiement, et ainsi de responsabiliser les personnes les mieux à même d’atténuer les risques

spécifiques. En outre, de nombreuses technologies linguistiques utilisées aujourd'hui s'appuient sur des modèles préexistants sur lesquels les utilisateurs n'ont que peu de visibilité.

Les modèles de langage d'IA présentent des risques pour les droits de la personne, la vie privée, l'équité, la robustesse, la sécurité et la sûreté.

Les modèles de langage d'IA sont une forme d'« IA générative ». Les modèles d'IA générative créent de nouveaux contenus en réponse à des invites, sur la base de leurs données d'apprentissage. Les données d'apprentissage elles-mêmes peuvent inclure des préjugés, des informations confidentielles sur des personnes et des informations couvertes par la protection de la propriété intellectuelle existante. Les modèles de langage peuvent alors discriminer, divulguer ou déduire des informations confidentielles ou qui portent atteinte aux droits des personnes.

Les modèles de langage d'IA peuvent aider à manipuler les opinions à grande échelle et à automatiser la désinformation d'une manière qui peut menacer les valeurs démocratiques.

Les modèles de langage d'IA peuvent faciliter et amplifier la production et la diffusion de fausses nouvelles et d'autres formes de contenu manipulé basé sur le langage, qu'il peut être impossible de distinguer d'informations factuelles, ce qui présente des risques pour la démocratie, la cohésion sociale et la confiance des citoyens dans les institutions. On assiste également à des « hallucinations » de l'IA, lorsque les modèles génèrent des résultats incorrects mais les expriment de manière convaincante. La combinaison des modèles de langage de l'IA et de la désinformation peut conduire à une tromperie à grande échelle que les approches traditionnelles telles que la vérification des faits, les outils de détection et l'éducation aux médias ne peuvent pas facilement corriger.

La poursuite du dialogue et de la recherche peut contribuer à atténuer les risques liés aux modèles linguistiques complexes.

Il est essentiel de poursuivre les recherches pour comprendre ces modèles complexes et trouver des solutions d'atténuation des risques. En outre, si l'on se tourne vers l'avenir, des mécanismes de sécurité pourraient être nécessaires pour contrôler certaines formes de modèles linguistiques puissants qui peuvent avoir une incidence directe sur le monde réel. Une question quant aux implications sociétales importantes demeure : les modèles de langage puissants devraient-ils pouvoir agir directement, comme envoyer des courriels, faire des achats et publier des messages sur les médias sociaux, alors qu'ils sont actuellement utilisés comme des systèmes passifs de réponse à des interrogations ?

Les ressources linguistiques et de calcul ainsi que les modèles linguistiques sont essentiels.

L'accès à des textes lisibles numériquement pour former des modèles reste un problème important pour de nombreuses langues moins utilisées. Les modèles linguistiques les plus avancés utilisent les langues pour lesquelles un contenu numérique important est disponible, comme l'anglais, le chinois, le français ou l'espagnol. Dans les pays où les langues sont minoritaires, les décideurs politiques encouragent le développement de dépôts, de plans et de recherches sur les langues numériques. Les modèles linguistiques multilingues peuvent favoriser l'inclusion et profiter à un plus grand nombre de personnes. L'accès au matériel informatique est également crucial, mais il nécessite davantage de R&D pour réduire les coûts financiers et environnementaux en faveur de mécanismes plus efficaces d'apprentissage et d'interrogation de modèles linguistiques de grande taille.

Pour que les avantages des technologies linguistiques soient largement partagés, les acteurs devront se préparer aux transitions économiques à venir et former les personnes à l'élaboration et à l'utilisation de modèles de langage d'IA.

Les modèles linguistiques ont le potentiel d'automatiser des tâches dans de nombreuses catégories professionnelles. Les modèles linguistiques d'IA tels que le *Generative Pre-trained Transformer 4* (GPT-4) de l'OpenAI sont de plus en plus utilisés pour aider à réaliser des tâches auparavant effectuées par des personnes, y compris des tâches hautement qualifiées telles que l'écriture de codes de logiciels, la rédaction de rapports et même la création de contenu. Les GPT-4 affichent des performances égales à celles des humains à des tests standardisés. Les décideurs politiques devront expérimenter de nouveaux modèles sociaux et réévaluer les besoins en matière d'éducation à l'ère de l'omniprésence des modèles de langage de l'IA.

Une coopération internationale, interdisciplinaire et multipartite pour des modèles de langage d'IA dignes de confiance est nécessaire pour lutter contre les utilisations abusives et les effets néfastes.

Les parties prenantes, y compris les décideurs politiques, commencent à explorer les impacts et les risques sociétaux connexes. La collaboration prend la forme d'un partage des meilleures pratiques et des enseignements tirés dans les forums régionaux et internationaux et d'initiatives conjointes en matière de données et de modèles linguistiques multilingues. Il reste cependant du travail à faire pour élaborer des solutions politiques et techniques viables permettant d'atténuer efficacement les risques liés aux modèles linguistiques et à d'autres types d'IA générative, tout en favorisant leur développement et leur adoption bénéfiques. Tous les acteurs de l'écosystème de l'IA ont un rôle clé à jouer.

1 National AI policies and initiatives for language models

Language models underpin natural language processing (NLP) that automates a variety of natural language functions

Natural language processing (NLP) refers to computer programs and tools that automate natural language functions by analysing, producing, modifying, or responding to human texts and speech. NLP is a subset of artificial intelligence (AI) that uses language as an input, produces language as an output, or both. Language models take centre stage in NLP. They are models designed to represent the language domain that often use machine learning. Chatbots, machine translation systems, and virtual assistants that recognise speech are all applications that use language models.

Policy makers are actively encouraging or guiding the development and deployment of NLP in national languages

National governments are recognising the growing importance of AI language models and other NLP applications to enhance public services, promote national languages, boost productivity and decrease costs. This section explores current and emerging policy initiatives, including national action plans and strategies, to encourage or guide the development and deployment of NLP in national languages.

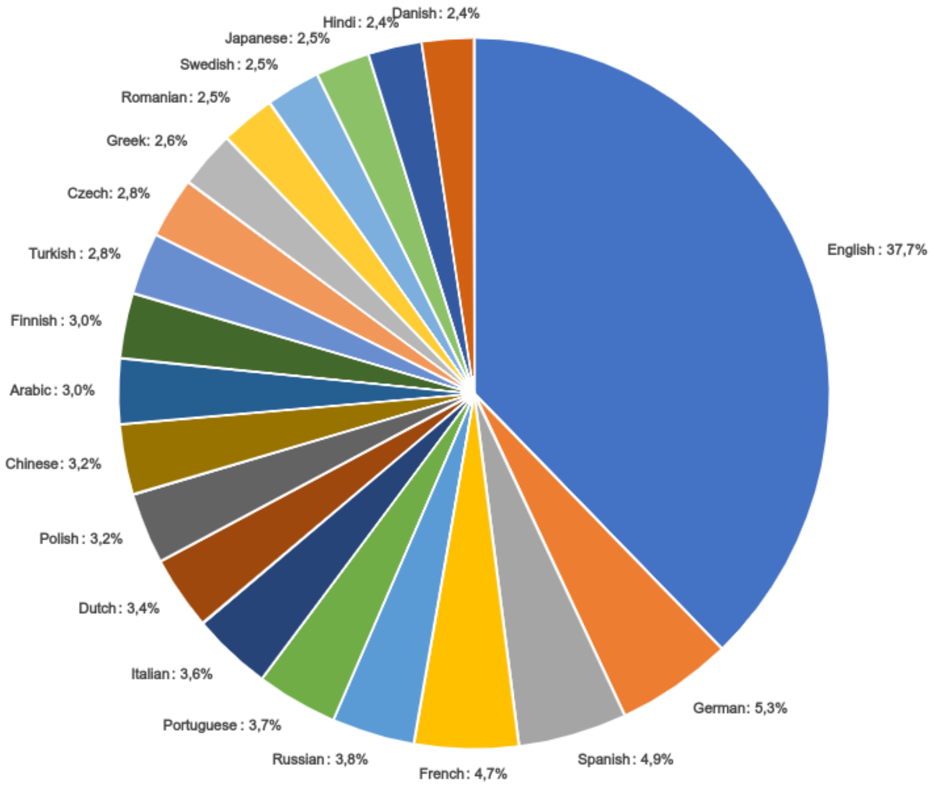
A key trend in this space is investment in developing digital language resources in non-English languages, including less commonly used languages or indigenous languages, particularly due to the lower availability of AI training data in languages other than English. According to Hugging Face, a repository of open-source NLP models, datasets, and libraries, resources in English represents 38% of all language resources, followed by Spanish, German, and French (Figure 1), while training data in Chinese is reportedly significant but less well reported on that specific platform. NLP research centres and collaborative platforms have been created with networks of partners from the private sector, academia, and civil society. A growing number of cross-border initiatives aim to share know-how and best practices and to facilitate the interoperability of national language data systems.

OECD countries

In **Canada**, the National Research Council's multilingual text processing team carries out R&D in multilingual AI language models, including machine translation and other NLP systems for multilingual contexts (National Research Council Canada, 2021[1]). Notably, the Canadian indigenous languages technology project provides AI language models in support of Indigenous language schools, educators, students, communities, and technology developers. Collaborative projects generate new speech- and text-based resources for these groups and help increase the accessibility of audio and video recordings.

In **Denmark** in 2021, the government committed to investing EUR 4 million in a Danish language resource as part of the Danish National Strategy for AI (Ministry of Foreign Affairs of Denmark, 2021[2]). Now, the Danish Gigaword Project assembles the first dataset with over one billion Danish words, which can increase the accuracy of automated translation services and other NLP applications. The University of Copenhagen's Centre for Language Technology research focuses on five themes: Language Processing and Resources; LT applications; machine learning; multi-modal communication which integrates text and other modes of communication; and language research infrastructure (University of Copenhagen, 2023[3]).

Figure 1. AI datasets on Hugging Face by language, 2023



Note: Top 20 languages from Hugging Face AI datasets, from a list of 225 languages. Given that the biggest portion of the datasets belongs to smaller languages, selecting the top 20 languages was needed for visualisation purposes.
Source: OECD.AI (2023), visualisations powered by Joseph Stefan Institute (JSI) using data from Hugging Face, www.oecd.ai/data.

In **Estonia**, the LT R&D programme, which is called "Estonian Language Technology 2018-2027," helps achieve the LT-related objectives for "Knowledge-based Estonia 2014-2020" and the "Estonian Language Strategy 2018-2027" (Estonian Ministry of Education and Research, 2018[4]). This programme aims to ensure that the essential components of Estonian language models comply with international standards and that Estonian language systems can be used more widely. The government also announced the development of a central translation platform with the overarching goal of increasing the digital accessibility of the Estonian language (Multilingual, 2021[5]).

In **France**, the government created the National AI Research Programme, which includes LT programs in four interdisciplinary AI institutes in Toulouse, Nice, Grenoble and Paris (Ministère de la Culture, 2021[6]). It developed the "PIAF" (Pour des IA Francophones) (Piaf, 2023[7]) project to build French-language datasets for AI in an open and contributory way (Ministère de la Culture, 2021[6]). PIAF will produce a dataset of 25 000 questions and answers in French. The government has a number of LT policy initiatives, including a national AI roadmap for AI language models and NLP systems, the opening of the Cité internationale de la langue française et de la Francophonie that should become an LT laboratory, and a proposed International Centre for Digital and LTs (Ministère de la Culture, 2021[6]).

In **Finland**, the Prime Minister's Office awarded Latvian company Tilde a four-year contract to provide machine translation services to the Finnish government and its ministries (Tilde, 2022[8]). It developed AURA, which generates translations of documents and texts for Finnish-Swedish, Swedish-Finnish, Finnish-English and English-Finnish. These custom machine translation engines are based on machine learning algorithms with integrated dynamic terminology for the unique terminology and linguistic requirements. There is a Finnish Government Termbank Valter, a multilingual databank containing glossaries compiled by the Terminology Service of the Prime Minister's Office (LT-Innovate, 2023[9]).

In **Germany**, the German version of the GPT-2 model has been trained on a massive German language dataset. As well, the German Society for Computational Linguistics and LT offers research, teaching and professional work in NLP (LT-Innovate, 2023[10]). The German Federal Ministry for Economic Affairs and Climate Action created the SPEAKER project to develop a "made-in-Germany" voice assistant platform. The German Research Centre for AI houses the German Competence Centre in Speech and Language Technology. The German Society for Computational Linguistics and Language Technology is the German-speaking countries and regions' scientific association for NLP research and teaching.

In **Hungary**, LT initiatives focus on linguistic resources carried out at the Research Institute for Linguistics of the Hungarian Academy of Sciences (LT-Innovate, 2023[11]). The Research Institute has also participated in several international projects aimed at adopting specific processes developed for Western European languages. The Hungarian National Corpus, a reference corpus of present-day Hungarian, which consists of 187.6 million words, exists thanks to the work of the Hungarian Language Offices and the Department of Corpus Linguistics (LT-Innovate, 2023[11]). The AI Coalition, founded by the Ministry of Innovation and Technology, has an NLP group. It works on training a smaller Hungarian version of the GPT-3 language model. The government is in talks to share this language model with the governments of Slovenia, Slovakia and Czech Republic to develop a collection of language corpuses.

In **Israel**, NLP is one of the main pillars of the National Programme for AI Infrastructure initiated in March 2021. The programme, which is promoted by several entities including TELEM (Israel's National Infrastructure Forum for Research and Development) includes an investment of USD 55 million in infrastructure for Hebrew and Arabic. This programme plans includes developing an extensive set of technological assets and with an open access provision for the R&D community to promote innovation in the academic, private and public sectors. The programme's deliverables are expected to include a rich set of datasets and corpora, state-of-the-art large-scale LMs for Hebrew and Arabic, Automatic Speech Recognition (ASR) modules, a Hebrew-Arabic bi-directional translation model and a set of pre-trained models for everyday NLP tasks such as document summarisation, named entity recognition, sentiment analysis and text-based question answering, and more.

In **Japan**, the National Institute of Information and Communications Technology (NICT) conducts research in a few LT areas. Under the Global Communication Plan 2025, the government discussed developing and using NLP resources (Cabinet Office of Japan, 2021[12]). Its objectives include promoting global and free communication, enhancing global business capability, realising an inclusive and multicultural society, and increasing Japan's visibility in the global arena. A notable project under the Plan is between the NICT, the Ministries of Internal Affairs and Communications and of Health, Labour and Welfare to further enhance a multilingual translation system for the labour field (Ministry of Internal Affairs and Communications of Japan, 2021[13]).

In **Korea**, the Government has released its National AI Initiative for NLP systems. It has been building Korean data since 2017 – which is difficult for companies, researchers, and individuals to build on their own due to time and cost constraints. It does this through surveys covering the private and public sectors. It has released 49 types of data, including text summaries of Korean language documents, voice data of Korean dialects and Korean-foreign language translation corpora. In 2022, the government was planned to release 44 additional types of Korean language data, including texts of Korean dialogue and multi-speaker voice datasets constructed in 2021.

In **Latvia**, the importance of AI language models, along with the need for digital language resources and language repositories, is highlighted in policy planning documents for 2021-2027, including the Digital Transformation Guidelines, the State Language Policy Guidelines and the report "On the development of AI solutions". State Research programmes such as "Digital resources for humanities: integration and development" and "Letonika" address AI LTs, focusing on creating and extending language resources. In Latvia's Recovery and Resilience Facility Plan, investments are allocated to developing and implementing high-level skills in AI language models (Viksna et al., 2022[14]).

The Latvian National Corpora Collection (LNCC) includes a vast collection of open-access text and speech corpora created by the digitalhumanities.lv community accessible through the korpuss.lv web platform. There has been substantial progress in digitising archive, library and museum collections, such as those of the National Library of Latvia (Periodika.lv) and the Latvian State Archives of Audio-visual Documents (redzidzirdilatviju.lv). The CLARIN (Common Language Resources and Technology Infrastructure) Centre of Latvian language resources and tools supports and collaborates with digital humanities and AI language model developers and provides sustainable access to Latvian language resources and tools in accordance with Open Data and FAIR Data Principles. Finally, the Latvian state administration has developed a platform known as Hugo.lv, which makes AI language models and services available to public administrations and the general public.

In **Lithuania**, the State Commission of the Lithuanian Language released the Guidelines for the Development of the Lithuanian Language in the Digital Environment and the Progress of NLP for 2021-2027 (Gaidienė and Tamulionienė, 2022[15]). The Guidelines aim to ensure the full functioning and use of the Lithuanian language in the digital environment and the progress of its "Lithuanisation," also to promote the development of technologies adapted to the Lithuanian language and to improve the quality of public services based on such technologies. Also, Latvian company Tilde has collaborated with Vilnius University to develop a free machine translation platform for Lithuanians in order to ensure information and service accessibility (Tilde, 2021[16]).

In **Norway**, the Norwegian Language Bank serves as a national infrastructure for NLP resources and offers datasets with Norwegian speech and text (UNESCO, 2020[17]); (National Library of Norway, 2023[18]). The Language Bank was established in 2010 with a yearly funding of NOK 10 million. In 2019, the government allocated an additional NOK 9 million to the project. The government is working with public- and private-sector partners on a few collaborative projects related to AI language models. For example, the Research Council of Norway is funding the SCRIBE project, whose mission is to design a Norwegian Speech-to-Text Transformer (S2T) transcription system for multi-party conversations in realistic recording conditions (SCRIBE, 2021[19]). The Council is funding CLARINO (CLARINO, 2020[20]), which is Norway's part of a broader EU initiative called CLARIN to promote greater access to language resources (CLARIN, 2022[21]).

Norway also plays a vital role in the Nordic Language Processing Laboratory (NLPL), a collaboration of university research groups in NLP in Northern Europe (Nordic Language Processing Laboratory, 2021[22]). The NLPL comprises NLP research groups in Norway, Finland, Denmark, and Sweden and the national e-infrastructure providers of Finland and Norway. Its vision is to implement a virtual laboratory for large-scale NLP research by (a) creating new ways to enable data- and compute-intensive NLP research by implementing a common software, data and service stack in Nordic High Performance Computing (HPC) centres; (b) pooling competencies within the user community and among expert support teams; and (c) enabling internationally competitive, data-intensive research and experimentation on a scale that would be difficult to sustain on commodity computing resources.

In **Slovenia**, the Research Infrastructure Development Plan 2011-2020 identified a need to set up an infrastructure for language resources and technologies (LT-Innovate, 2023[23]). This plan includes CLARIN (Common Language Resources and Technology Infrastructure) as a priority research infrastructure. Three departments at the Jožef Stefan Institute (JSI), a key technical partner of the OECD.AI Policy Observatory, are involved in AI language model research for both Slovenian and English: AI Laboratory, Department of Knowledge Technologies, Department of Intelligent Systems (LT-Innovate, 2023[23]).

In **Spain**, the government funded a National Plan for the Advancement of Natural Language Technologies (OECD.AI, 2021[24]). In December 2020, Spain presented its National AI Strategy (ENIA) to guide the country's AI plans and strategies over the period 2020-2025. ENIA's measures include the promotion of a new National Language Technologies Plan. In 2022, the government launched its New Language Economy project to mobilise public and private investments to promote Spanish and the country's co-official languages Catalan, Valencian, Galician, Basque, and Aranese (Kückens, 2022[25]). A total of EUR 1.1 billion is planned to be invested in the initiative by early 2026.

Spain's co-official languages are supported by regional government initiatives, including the ongoing AINA project generating new NLP resources and annotated datasets for Catalan. Similarly, NÓS, GAITU and VIVES respectively address Galician, Basque and Valencian, which currently have a significantly lower number of linguistic resources. The Spanish Secretary of State for Digitisation and Artificial Intelligence co-funds these projects and promotes other initiatives, such as the Spanish Language and Artificial Intelligence (LEIA) project led by the Spanish Royal Academy.

The **Republic of Türkiye** (hereafter 'Türkiye') launched its "National Artificial Intelligence Strategy 2021-2025" in 2021, including a competition and support programme to promote the emergence of at least one global Turkish NLP actor. The "2023 Industry and Technology Strategy" established "Open-Source Platform Türkiye" to host open source projects, including NLP projects, to strengthen the country's open-source ecosystem, upskill software developers and develop Turkish language models. The Ministry of Industry and Technology also developed an NLP Project to make available high-performance libraries and datasets required for the processing of Turkish language text.

In the **United Kingdom**, Wales, the Welsh Language Technology Action Plan was launched in 2018 out of the Welsh government's language strategy "Cymraeg 2050: A million Welsh speakers" (OECD.AI, 2021[26]). The Action Plan aims to plan technological developments to ensure that the Welsh language can be used in different contexts, whether through voice, keyboard, or other means of human-computer interaction. Its focus areas include Welsh Language Speech Technology, Computer-Assisted Translation, and Conversational AI.

Partner economies

In **Brazil**, the Brazilian Computer Society organised regular meetings called the "Workshop on Information Technology and Human Language" (STIL, in Portuguese) to bring together researchers, academics and businesses working on computing and linguistics. STIL has become the central NLP forum in Brazil and brings together multidisciplinary communities to advance the computational processing of human languages.

In the **People's Republic of China** (hereafter 'China'), the government-funded Beijing Academy of AI (BAAI) announced the first two versions of its Wu Dao LM in 2021. It has been reported that Wu Dao 2.0 contains more than 1.75 trillion parameters, which surpasses the number of parameters contained by Google's Switch Transformer, OpenAI's GPT-3 and Google's GLaM (Romero, 2021[27]). BAAI has partnered with a total of 22 Chinese companies, including smartphone company Xiaomi, delivery service company Meituan and media company Kuaishou.

In **Egypt**, the Ministry of Communication and Information Technology is engaged in a three-year agreement with Chinese company iFlytek to encourage R&D in NLP and machine translation in the Arabic and Chinese languages (Marking, 2020[28]). The Ministry's Applied Innovation Centre and iFlytek are creating research projects in Arabic speech recognition, speech synthesis and Chinese-Arabic translation (iFlytek, 2020[29]). In addition to technical co-operation, the two parties are co-ordinating the programme's implementation and exchanging information.

In **India**, the Ministry of Electronics and Information Technology (MeitY) established the Technology Development for Indian Languages (TDIL) Programme (Government of India, 2021[30]). The objective is to create information processing tools to facilitate human-machine interaction in Indian languages and develop technologies to access multilingual knowledge resources (Government of India, 2021[30]). The Ministry is also working on a technology to translate vernacular Indian languages in real-time to enable the exchange of communications between two persons not speaking the same language (Press Trust of India, 2021[31]). The government announced the National Language Translation Mission, which uses NLP and NLU to make governance- and policy-related knowledge available and to make knowledge texts accessible in all 22 official languages of India while enhancing access to digital content and online services for Indian users (Digwatch, 2021[32]). The National Language Translation Mission focuses on local language translation to enhance access to science and technology-related content in local communities via digital platforms. The AI4Bharat (AI4Bahrat, 2022[33]) project of the Madras Indian Institute of Technology aims to "Bring parity with respect to English in AI technologies for Indian languages with open-source contributions in datasets, models, and applications and by enabling an innovation ecosystem".

In **Qatar**, the Arabic Language Technology Group at the Qatar Computing Research Institute conducts research in NLP areas, including speech recognition, machine translation and question answering for the Arabic language (Arabic Language Technology Group, 2022[34]). It collaborates with local and international organisations, such as Al Jazeera, MIT CSAIL, and the Qatar Supreme Education Council.

In **South Africa**, the South African Centre for Digital Language Resources (SADiLaR) is a platform, which creates and manages digital resources and software to support NLP R&D and related studies in the nation's eleven official languages (Government of South Africa, 2019[35]). It is a collaborative infrastructure with a vast network of partner organisations, including the Council for Scientific and Industrial Research, the Universities of South Africa and Pretoria, the Inter-institutional Centre for Language Development and Assessment and North-West University's Centre for Text Technology. SADiLaR is hosted by North-West University at the Potchefstroom Campus. The Masakhane NLP organisation (Masakhane, 2020[36]) is a grassroots organisation whose mission is "to strengthen and spur NLP research in African languages, for Africans, by Africans".

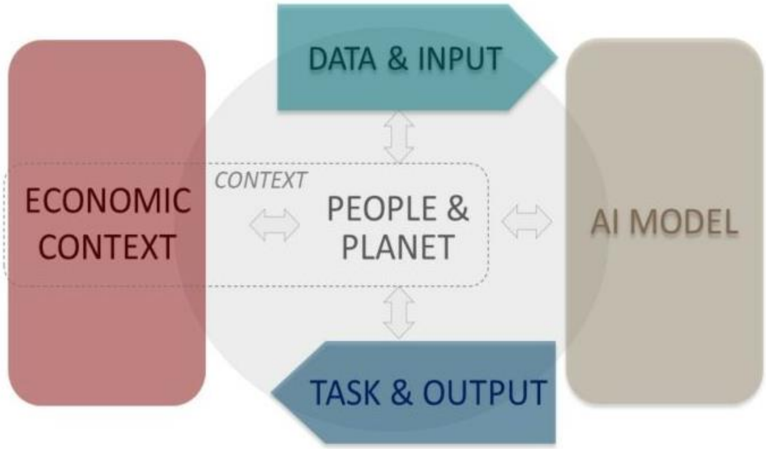
In **Thailand**, the National Electronics and Computer Technology Centre designed the AI FOR THAI platform as a digital infrastructure for the Thai people (Tapsai, Unger and Meesad, 2021[37]). The platform is in the form of API services for users and developers to create and develop applications to benefit businesses and society. AI FOR THAI offers a total of eleven services related to language technologies, including basic NLP; tag suggestion; machine translation; sentiment analysis; character recognition; object recognition; face analytics; persons and activity analytics; speech-to-text; text-to-speech transformers (T2S); and chatbots.

In **Viet Nam**, the Ministry of Information and Communications developed an AI-based Vietnamese-language S2T generator called VAIS and a T2S engine called Vbee (Dharmaraj, 2020[38]). AIS can recognise Vietnamese accents from all northern, central and southern regions with an accuracy rate of up to 95 per cent and immediately produce results. It has been used by the Offices of the Party Central Committee, the Government and the National Assembly, the Ministry of Information and Communications and the Hanoi People's Committee.

2 Understanding the evolving ecosystem for AI language models

This section examines the technological dimensions of AI language technologies through the lens of the OECD Framework for the Classification of AI Systems – namely, how they impact people and planet, the economic context in which systems operate, the data and input i.e. language resources used, AI models and the tasks and outputs of the systems. It provides an overview of the technical components of language models and stakeholders. Although the OECD Framework for the Classification of AI Systems is primarily designed to classify applied AI systems, this report also applies the Framework to general-purpose AI language models that are then integrated into specific applications. Notably, the general-purpose Large Language Model (LLM) GPT-3 serves as a use case (Annex B).

Figure 2. Five dimensions of the OECD Framework for the Classification of AI Systems



Source: (OECD, 2022[39])

People and planet

Impacted stakeholders include the general public, workers, consumers, and those without access to LMs

The people and planet dimension is at the centre of the OECD Framework and focuses on "… human rights and well-being in considering how people as a whole interact with and are affected by an AI system throughout that system's lifecycle" (OECD, 2022[39]). Users of AI language models and applications include the public, who may lack AI competency, businesses, research institutions and government agencies. In practice, they are currently skewed towards English speakers, as many existing language resources and models are in English. Within this broad category of users, key impacted stakeholders include workers and consumers, as AI language models could lead to the automation of some tasks, such as translation and question-answering. Another group of impacted stakeholders includes those non-users who may not have access to AI language models and in turn are at a disadvantage in unlocking the economic benefits and opportunities that these systems can provide.

Economic context

The economic impact of AI language models is already considerable and poised to be transformational

The economic context "constitutes the economic and sectoral environment in which an AI system is deployed" (OECD, 2022[39]). This includes the sector for which it is developed and deployed, business function and model, criticality and its scale and technological maturity. Actors in AI language models include system operators and users, such as large technology companies and AI research organisations and firms involved in studying computational linguistics. Several technology companies are researching, designing and deploying general-purpose NLP systems on a large scale, especially in the United States and China (e.g. ChatGPT by OpenAI, Claude by Anthropic, Bard by Google, etc.). They have applications in many sectors including public administration, healthcare, law, and retail.

The economic impact of language models is considerable and continuing to grow, particularly considering the rise of large AI language models. Language, in written, spoken, or visual-manual form, is the main vector of human communication as it helps people to function in society, teaching them how to socialise and learn (see also Annex A). Language skills can provide economic benefits to individuals and societies, enhance intercultural skills and global co-operation and lead to new and innovative ways of thinking and working across cultures (Marconi et al., 2020[40]). By extension, AI language models can significantly impact both industry and government.

AI language models are already starting to impact approaches to learning and using language – for example, through applications like Duolingo, Busuu and Grammarly. Language models provide innovative solutions to break some of the language barriers associated with transnational communication in areas such as international trade and are being used to optimise the efficiency of operations in sectors such as healthcare, commerce, banking, insurance, and public administration (Dilmegani, 2022[41]).

Data and input

Data and input include language resources, grammar and terminology databases, annotated corpora and human feedback

The data and input dimension is defined as "the data and/or expert input with which an AI model creates an artificial representation of the physical environment" (OECD, 2022[39]). In the LM context, the data refers to language resources, including written and spoken corpora (such as CommonCrawl and WebText) and grammar and terminology databases (Box 1). In addition to raw data for unsupervised language models, another critical piece of the data is supervised training data, usually in the form of annotated corpora. Some modern AI language models such as GPT-4, are also trained on human feedback provided in response to generated outputs. While most existing language resources are in English, this is changing, as non-English-speaking countries recognise the importance of language data and developing language resources in their own languages.

Box 1. Language resources in the context of the European Language Grid and the European Language Equality initiatives

The European Language Grid (ELG) and the European Language Equality (ELE) are related projects that focus on closing the digital language divide. ELG is an online platform which catalogues European language technologies, language resources and members of the European language community. ELE's areas of focus include Machine Translation, Speech Technologies, Text Analytics and Natural Language Understanding and Data.

In the *Report on the State of the Art in Language Technology and Language-centric AI* (Rodrigo et al., 2021[42]), ELE outlined catalogues where European language resources can be found, including the European Commission's ELRC-SHARE, the European Language Grid, the European Language Resources Association (ELRA) Catalogue, the Common Language Resources and Technology Infrastructure (CLARIN) and META-SHARE. Beyond Europe, ELE refers to the Linguistic Data Consortium as both a repository and a distribution point for language resources. ELE notes that even though there are notable language resources in English, French, German, Spanish and Italian, to date English remains far ahead of the rest.

Source: (European Language Grid, 2023[43]; European Language Equality, 2023[44]).

Data selection and curation are essential for trustworthy AI language models

The selection and curation of data is essential for the development and deployment of trustworthy AI language models. It is important for model developers to consider the representativeness or accuracy of the data in datasets since models can reproduce inaccurate or incorrect information present in the training data. It is also essential to assess whether the data contains private information or biases (OECD, 2022[39]). Selecting datasets that consider these factors provides a solid foundation for training language models. In addition to selecting reliable and accurate datasets, it is important to curate selected datasets, i.e. label and clean data for inconsistencies and remove personal information and biases where necessary.

AI language model

AI language models are often characterised by their parameter count and layers and accuracy

The OECD Framework for the Classification of AI Systems defines an AI model as "a computational representation that encompasses processes, objects, ideas, people and interactions. Language models vary in language and size. AI language models are often characterised by their parameter count and layers and accuracy. A parameter refers to a value that the model changes independently through training for a specific task (Wiggers, 2022[45]). Parameters are distributed and applied across layers, which refer to various stages of processing or iteration for an input sequence (Simoulin and Crabbé, 2021[46]).

Progress in AI LMs since 2015 has been notable, thanks to neural networks, large datasets and compute

While NLP has existed since the 1940s, significant strides in research and development have been made over the past few years. Progress since 2015 has been notable, thanks to machine learning neural network models, alongside the availability of large volumes of data and computational power. Neural networks have enabled state-of-the-art results in natural language tasks and notably in the development of language models. For example, the "Google Translate" service, which switched from statistical machine translation to neural machine translation in 2016, uses machine learning and deep neural network models trained on large volumes of full sentence translations of various official data sources of such as United Nations and European Parliament documents (Ye, 2020[47]). Rather than individual words, it is trained on full sentences, dramatically improving quality and accuracy and enabling new functionalities.

The 2017 "Transformer" architecture was a critical conceptual breakthrough that enabled powerful LLMs

The "Transformer" architecture developed in 2017 was the critical conceptual breakthrough that brought further significant progress to language models. It is more efficient than its predecessor, Recurrent Neural Networks (RNN), because Transformers can process natural language input in parallel rather than merely in sequence. This reduces training and computing time and allows the development and deployment of Large Language Models (LLMs) (Vaswani et al., 2017[48]). The Transformer's architecture comprises three technical concepts: positional encodings, attention and self-attention (Bilogur, 2020[49]). "Positional encoding" provides information on the position of each part of an input sequence, e.g. a sequence of text. "Attention" draws connections between the different parts of the sequence, allowing a language model to focus on previously hidden vectors in an input sequence to predict an output sequence. "Self-attention" allows both previously hidden vectors and later parts of the input sequence to interact with one another (Bilogur, 2020[49]). Actors developing Large Language Models (LLMs) include OpenAI, Google and Deepmind, Meta, as well as Baidu, the Beijing Academy of AI (BAAI), Amazon, Hugging Face, Yandex and Anthropic (Box 2).

Box 2. Notable Large Language Models (LLMs)

Many LLMs have been released by companies and other organisations in recent years.

OpenAI

In 2020, OpenAI unveiled the Generative Pre-Trained Transformer 1 (GPT-1) in 2018 and subsequent versions GPT-2 in 2019 and GPT-3. GPT-3 is a pre-trained LLM which upon receiving a text prompt, GPT-3 is able to return a text completion in natural language.

In 2022, it released a conversational AI language model called ChatGPT using an enhanced GPT-3.5. According to OpenAI, the dialogue format of ChatGPT enables it to answer follow-up questions, acknowledge its mistakes, challenge incorrect premises and reject inappropriate requests. ChatGPT was trained using "Reinforcement Learning from Human Feedback", which involves using methods from reinforcement learning to directly optimise an LM with human feedback (Hugging Face, 2022[50]).

In 2023, OpenAI launched a multi-modal LLM called GPT-4, which accepts image and text inputs and emits text outputs. GPT-4 is now used on ChatGPT for some users. According to the company, the post-training alignment process results in improved performance on measures of factuality and adherence to desired behaviour (OpenAI, 2023[51]).

Google and Deepmind

In 2018, Google launched its "Bi-directional Encoder Representations from Transformers" (BERT) (Devlin and Chang, 2018[52]). BERT was the first bi-directional and unsupervised language representation to be pre-trained on a plain text corpus (Devlin and Chang, 2018[52]).

In 2021, Google launched its Switch Transformer and the Generalist Language Model (GLaM), which allows the reduction of computational and financial costs by reducing the number of "experts" (i.e. submodels) activated (e.g. for each query) (Google, 2021[53]).

Also in 2021, Google's subsidiary Deepmind unveiled Gopher, a Transformer-based AI language model, with 280 billion parameters. It outperformed many other models at the time, especially in knowledge-intensive domains.

In 2022, the company put out its 540-billion-parameter Pathways Language Model (PaLM), which outperformed similar LLMs such as GLaM, GPT-3, Megatron-Turing NLG and Gopher in few-shot performance (i.e. its ability to generalise information using few examples). The same year, Deepmind introduced its 70-billion-parameter AI language model called Chinchilla, which is reported to outperform Gopher and GPT-3 on a range of downstream evaluation tasks.

Meta

In 2022, Meta launched its Open Pre-Trained Transformer (OPT-175B), which is the first 175-billlion-parameter LLM made available to the broader AI research community (Meta, 2022[54]). The release includes both the pretrained models and the code needed to train and use them. Its performance is estimated to be comparable to GPT-3, but with only one-seventh the carbon footprint (Synced, 2022[55]).

In 2023, Meta released LLaMA, a foundational LLM designed to help researchers advance their work in AI by providing a smaller LLM that does not require major infrastructure to operate (Meta, 2023[56]).

Others

Baidu. In 2019, Chinese technology giant Baidu launched the first version of its AI language model ERNIE ("Enhanced Representation through kNowledge IntEgration"), which was followed by ERNIE 2.0 that same year and ERNIE 3.0 in 2021 (Synced, 2021[57]). According to Baidu, the training data for ERNIE 3.0 is the largest Chinese language dataset to date.

The Beijing Academy of AI (BAAI). In 2022, BAAI announced the first two versions of its Wu Dao LLM, which contains more parameters than GPT-3 (Rodriguez, 2021[58]).

Amazon. In 2022, Amazon released its 20-billion-parameter language model called the Alexa Teacher Model 20B (AlexaTM 20B) (Amazon, 2022[59]). It outperformed some larger models in few-shot learning and requires lower carbon consumption in its training (Gupta, 2022[60]). Amazon has also applied ML techniques in an effort to debias and detoxify the model (Amazon, 2022[59]).

Hugging Face. In 2022, Hugging Face released its BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) in co-operation with several French governmental agencies (Hugging Face, 2022[61]). BLOOM includes 176 billion parameters and was in complete transparency on France's Jean Zay public supercomputer (Hugging Face, 2022[61]). Hugging Face has stated that BLOOM is the first multilingual open LLM whose training and deployment has been made transparent to the broader community.

Yandex. In 2022, the Russian firm published its 100-billion-parameter open-source LM called YaLM 100B, which the company describes as a bilingual neural network for generating and processing text (Yandex, 2022[62]).

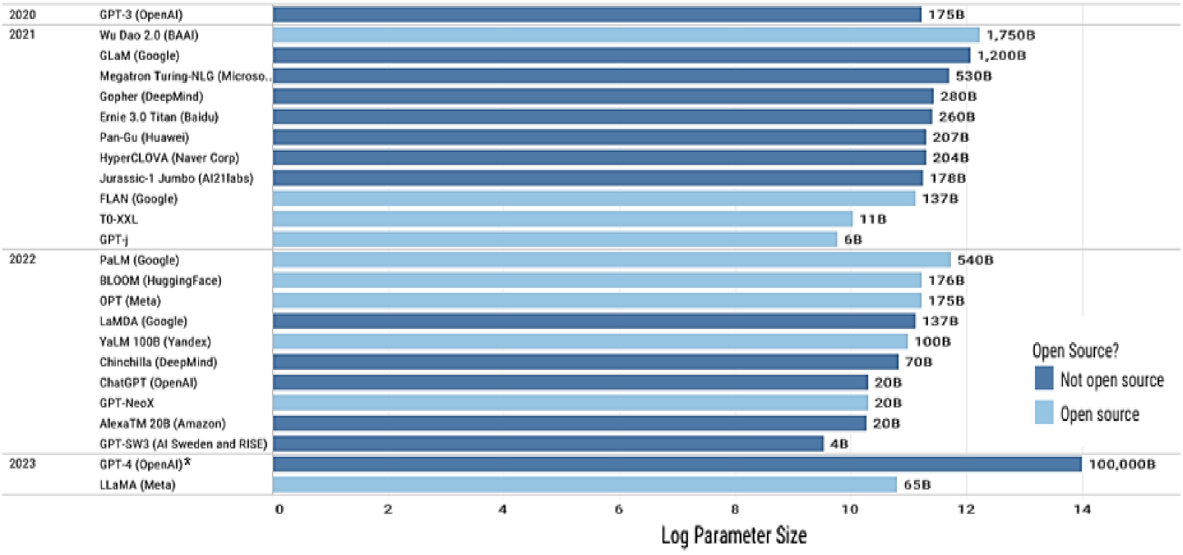
Anthropic. In March 2023, the company launched Claude, an "AI assistant based on Anthropic's research into training helpful, honest, and harmless AI systems" (Anthropic, 2023[63]).

Parameter count has been central to AI LM research, but other factors such as training efficiency matter

Parameter count is at the centre of research for AI language models. With the Transformer architecture, AI developers were able to design larger models with an exponential rise in the number of parameters (Figure 3). A higher number of parameters enables the AI language model to be more comprehensive in nature and in turn, in model performance. However, increasing the number of parameters may not always translate into an equivalent increase in performance within large language models and diminishing marginal returns may come into play at some point (Simon, 2021[64]).

Figure 3. Large language models are getting very big, and many are open source

The evolution of select large language models and their parameter counts (billions), 2020-2023



Note: This figure samples large language models and is not exhaustive. *The number of parameters for GPT-4 is an estimate by external sources and at the time of writing has not been disclosed by OpenAI.
Source: Based on research from OECD, (Sevilla et al., 2022[65]; Benaich et al., 2022[66]).

Researchers are exploring whether transformers can allow the creation of a generalist AI model

Some researchers are exploring whether transformers can allow the creation of a "generalist" AI model capable of performing hundreds of different tasks. While some researchers suggest such advancements pave the way for human- or superhuman level intelligence, known as Artificial General Intelligence (AGI) (Box 3), these models remain applicable to relatively narrow contexts, exhibiting imperfections like "hallucinations" (making up facts if a right answer is not found in the training data), and often requiring human assistance and oversight for correct functioning.

Box 3. What is Artificial General Intelligence (AGI)?

With Artificial General Intelligence (AGI), autonomous machines would become capable of general intelligent action. Like humans, they would generalise and abstract learning across different cognitive functions. AGI would have a strong associative memory and be capable of judgment and decision making. It could solve multifaceted problems, learn through reading or experience, create concepts, perceive the world and itself, invent and be creative, react to the unexpected in complex environments and anticipate.

Despite significant technological advances over the last two decades, including the emergence of LLMs and generative AI, narrow AI largely remains the current state-of-the-art. However, the impressive performance of some recently released AI models have prompted claims of progress towards AGI. For example, researchers at Microsoft (Bubeck et al., 2023[67]) believe that OpenAI's GPT-4 could "reasonably be viewed as an early (yet still incomplete) version of an [AGI] system", exhibiting some elements of a theory of mind - the capacity to understand people by ascribing mental states to them (Wellman, 1992[68]). Although such AI models can generalise pattern recognition by transferring learning between text, speech, image, and video recognition with degrees of precision, they can still be prone to factual inaccuracy, hallucinations (making up facts if a right answer is not found in the training data), inconsistency, and misunderstanding in new contexts.

With respect to a potential AGI, views vary widely. Experts caution that discussions should be realistic in terms of time scales. They broadly agree that narrow AI will generate significant new opportunities, risks, and challenges. They also agree that the possible advent of an AGI, perhaps sometime during the 21st century, would greatly amplify these consequences.

Sources: (Russell and Norvig, 2016[69]; OECD, 2019[70]) (Bubeck et al., 2023[67])

To date, higher parameter counts have been viewed as automatically improving model performance

To date, language models trained on small volumes of language data have tended to work less well than models trained using large volumes of data (Paullada, 2021[71]). There has been a perception that the higher the parameter count the better the model performance. However, this attitude has started to shift among developers because: 1) the sheer scale of AI language models has raised ethical, environmental and security challenges that have prompted the development of smaller language models and 2) higher performance can also be achieved by optimising the efficiency of the training (Box 4). Compared with large language models, small language models tend to consume less computing power and focus more closely on specific operational needs. Yet depending on the specific language function(s) that one is looking to perform, less can be more. Small language models have lower hardware requirements and lower environmental and financial cost. At the same time, small language models may require significant R&D investment to achieve similar performance to large language models.

Box 4. Training neural networks to be more efficient

The OECD report *Measuring the environmental impacts of AI compute and applications: The AI Footprint* describes several good practices for sustainable AI, including the use of pre-trained models and powering data centres with renewable resources. For example, researchers involved in a project at Massachusetts Institute of Technology (MIT) and start-up MosaicML are training neural networks up to seven times faster by configuring AI algorithms to learn more efficiently. Efficiency gains for both computing hardware and software can be explored to maximise positive sustainability impacts in training and using AI systems. The goal for future machine learning models is to balance performance and efficiency as demand for AI computing grows with the creation of bigger and more complex AI language models.

Source: (OECD, 2022[72])

Small language models that use methods to make training more efficient hold significant promise

The immense resources required to develop and execute large models have raised ethical, environmental, and digital security challenges, which will be further discussed in the following section on policy considerations. Research institutions and large technology firms have thus also been developing leading small language models (Table 1). Higher parameter count and model size are not the sole performance indicators for language models. Other key factors include the size of the training dataset and the amount of compute used in training (Kaplan et al., 2020[73]). Optimal model performance involves finding an appropriate balance among these factors (Hoffman et al., 2022[74]). A growing number of researchers acknowledge that the size of an AI language model is not all that matters in optimising model performance. They illustrate that performance similar to GPT-3 can be achieved with smaller language models such as ALBERT, using additional training methods to make training more efficient (Schick and Schütze, 2021[75]).

Table 1. Selection SLMs sorted by year and by number of parameters

Model	Year of release	Developer (country)	Transformer type	Number of parameters (Millions)
ALBERT	2019	Google (US)	Autoencoding	31
BART	2019	Facebook (US)	Seq2seq	406
KoGPT	2021	Kakao (Korea)	Autoregressive	6 000
RETRO	2021	DeepMind (US)	Retrieval-Based	8 000
ERNIE 3.0	2021	Baidu (China)	Autoregressive; Autoencoding	10 000
T0	2021	BigScience (US)	Seq2seq	11 000

Sources: (Sun et al., 2021[76]; Borgeaud et al., 2022[77]; Lee, 2021[78]; Hugging Face, 2023[79]; Hugging Face, 2023[80]; Hugging Face, 2023[81]).

There is also growing interest in developing multilingual AI language models

Multilingual language models have been relatively limited in number because, until recently, much of the focus has been placed on monolinguafl or bilingual models. However, there is growing interest in developing multilingual AI language models, such as M-BERT by Google and XLM-R by Meta (Moberg, 2020[82]). Even though multilingual language models carry breadth in the number of languages they can detect and process, they can be less accurate for specific language pairs. This is because the amount of training data can vary by language and the model's different layers may require more language-specific information (Pires, Schlinger and Garrette, 2019[83]).

However, monolingual and bilingual language models are still very much in demand

Consequently, monolingual and bilingual language models are still very much in demand. As evidenced by the examples of leading language models, many are designed with and for high-resource languages such as English and Chinese due to the considerable amount of computing resources and training data required to develop and deploy them. Larger countries subsequently create their own versions of these language models in their native languages (Barba, 2020[84]). The next natural progression is that single multilingual language models are produced for the next one hundred most popular languages. This observation is reinforced in a 2021 European Language Equality (ELE) report on existing strategic documents and projects in NLP systems and AI: after English, only a handful of Western European languages dominate the field – in particular, French, Spanish and German – and even fewer non-Indo-European languages (European Language Equality, 2021[85]).

Offshoots of leading English language models have been adapted in other languages

As a temporary solution to the systemic issue of fewer AI language models in non-English-speaking countries, several offshoots of leading English language models have been adapted in other languages. For example, Facebook AI released the French version of BERT called CamemBERT. When compared to that of multilingual models in a number of downstream tasks, CamemBERT was superior in most (Martin et al., 2020[86]). Most recently, FlauBERT was also launched, which outperformed CamemBERT with all neural network architectures used for intent and slot prediction (Blanc. et al., 2022[87]). There are Spanish versions, including BETO, BERTIN and MarIA, and Norwegian versions, including NB-BERT and NorBERT (Rojas et al., 2020[88]; Hugging Face, 2021[89]; Kummervold et al., 2021[90]; Kutuzov et al., 2021[91]).

Language-specific models can be extended to other languages

Language-specific models can be extended to other languages through cross-lingual transfer learning, which involves leveraging labelled data from other source languages (Chen et al., 2019[92]). In effect, this method allows for the development of AI language models for low-resource languages through training data from higher-resource languages.

AI language models are the closest that has been achieved to date to what has been widely referred to in the AI community as generative AI. They can process immense amounts of natural language input and generate natural language output – and visual output in some cases, such as DALL-E. The Transformer's architecture which underlies many of the recent leading AI language models constitute a revolutionary development in NLP with tremendous potential for the global economy and society more broadly. At the same time, if not regulated properly, it is important to recognise that their capabilities can pose a significant risk to fairness, privacy, and security.

Task and output

The final dimension – task and output – includes the tasks performed by the AI system, its outputs, and resulting action(s) that influence the overall context. This section provides an overview of various NLP categories and their corresponding tasks.

Tasks include name entity recognition, part-of-speech tagging, text categorisation, syntactic parsing, and machine translation

AI language models can process text as input and/or generate text as output. NLP tasks include name entity recognition, part-of-speech tagging, text categorisation, syntactic parsing, and machine translation (Koradiya, 2019[93]; Dilmegani, 2022[41]). "Name entity recognition" involves persons, locations, and companies as the named entities. "Part-of-speech tagging" means that parts of speech are assigned to an input sequence. With "text categorisation", text in an input sequence is classified into pre-defined groups. With "syntactic parsing", an input sequence is broken down by grammatical structure. Finally, "machine translation" involves converting text in an input sequence from one language to another.

Natural language comprehension involves making natural language machine-understandable

The first subcategory of NLP involves natural language comprehension. It involves converting natural language input into a machine-understandable form. Examples of natural language understanding tasks include relationship extraction, semantic parsing, sentiment analysis and text summarisation (Koradiya, 2019[93]; Dilmegani, 2022[41]). "Relationship extraction" extracts semantic relationships between texts in an input sequence. "Semantic parsing" breaks down an input sequence into formal representations of its meaning, which can involve identifying the role of each piece of an input sequence, e.g. the agent, the task and the theme (Raj, 2017[94]). "Question answering" generates a response for a natural language input question. "Sentiment analysis" extracts emotional meaning from an input sequence. Finally, "text summarisation" shortens an input sequence without compromising its meaning.

Natural language generation refers to producing natural language, e.g. natural language conversation

A second subcategory of NLP is natural language generation, which refers to natural language production. Examples of Natural Language Generation tasks include discourse generation, lexical choice, sentence generation and document structuring (Koradiya, 2019[93]). "Discourse generation" generates natural language conversation between an NLP and a human agent. "Lexical choice" selects specific words in response to an input sequence, including generating a specific output word more suitable than others (Pandey, 2022[95]). For example, this involves mapping a concept (a person whose sex is male and whose age is between 13 and 15 years) to a word (a boy, kid, teenager, youth, child, young man, or schoolboy). Sentence generation means aggregating sentences in context by virtue of their relevance to an input sequence (Pandey, 2022[95]). For example, the input "Sam has high blood pressure. He has low blood sugar." would generate the output "Sam has high blood pressure and low blood sugar." "Document structuring" generates a narrative structure in response to an input sequence (Pandey, 2022[95]).

Automatic speech recognition refers to processing speech data as input, e.g. recognising the speaker

Another concept associated with NLP systems is automatic speech recognition, which refers to processing speech data as input. Common processes for improving Automatic Speech Recognition models for a specific application include language weighting, speaker labelling, acoustics training and profanity filtering (IBM, 2020[96]). “Language weighting” is when words spoken frequently are given a higher weight. “Speaker labelling” recognises speakers and cites or tags their contributions to a conversation. With “Acoustics training”, a system adapts to an acoustic environment –for example a level and type of background noise– and speaker styles –for example, certain accents–. Finally, “profanity filtering” identifies and filters out certain words or phrases and sanitise speech output.

3 Policy considerations

A growing number of policy challenges and considerations are associated with the adoption of AI language models and applications by individuals and organisations. This section leverages the 2019 OECD AI Principles to structure a discussion of salient policy considerations associated with AI language models. The OECD AI Principles are composed of five values-based Principles and five policy recommendations.

Research efforts also exist to categorise the risks to people and planet associated with AI language models (Box 5).

Box 5. Research on risks of harm to people and planet associated with AI language models

Researchers at DeepMind noted the following six risk areas to people and planet associated with language models:

- Discrimination, exclusion and toxicity: Harms that arise from the AI language model producing discriminatory and exclusionary speech.
- Information hazards: Harms that arise from the AI language model leaking or inferring true sensitive information.
- Misinformation harms: Harms that arise from the AI language model producing false or misleading information. Misinformation is shared unknowingly and is not intended to deliberately deceive, manipulate, or inflict harm on a person, social group, organisation or country (Lesher, Pawelec and Desai, 2022[97]).
- Malicious uses: Harms that arise from actors using the AI language model to intentionally cause harm.
- Human-computer interaction harms: Harms that arise from users overly trusting the AI language model or treating it as human-like.
- Automation, access and environmental harms: Harms that arise from the AI language model's environmental or downstream economic impacts.

Source: (Weidinger et al., 2021[98]).

Benefit people and planet (Principle 1.1)

AI language models can unlock tremendous opportunities across economies and societies

Natural language has long affected how people live and work. By extension, AI language models can make a positive difference in many contexts and sectors for inclusive growth, sustainable development, and well-being. AI language models are beginning to unlock tremendous opportunities by conducting tasks in human natural language on a large scale. They are being deployed across sectors, such as in public administration, healthcare, banking and law. They enable language recognition, interaction support and personalisation. Major NLP tasks include natural language understanding and generation and automatic speech recognition. For example, AI language models enable interactive dialogue systems and personal virtual assistants. In addition, LMs unlock socio-economic opportunities by decreasing language barriers on a large scale and by automating some language tasks.

While multi-lingual models help, limited training data remains an issue for many languages

However, populations without access to NLP applications in their dominant language miss opportunities, which is one reason that developing NLP models and applications in languages other than those that are widely used in AI-leading countries is a priority for many governments, academic and industry organisations. Recently, BigScience unveiled a multilingual open LLM called BLOOM that is available in 46 languages (Hugging Face, 2022[61]). Meta AI open-sourced NLLB-200, an AI model that can translate over 200 languages. A first step is generating sufficient digitalised natural language data to train language models. Otherwise, improvements in AI language models risk exacerbating the prominence of some languages (Borgonovi, Hervé and Seitz, 2023[99]).

Environmental concerns

The environmental and financial costs of developing and deploying AI LMs are significant

With all other factors remaining equal, as AI models grow in size and complexity (such as with LLMs), so do their computing demands. This results in environmental impacts including energy use, water consumption and CO2 emissions. Although researchers have tried to quantify the precise environmental impacts of AI language models, particularly for CO2 emissions, estimates vary and measurement methodologies and indicators could be further standardised to allow for comparability, benchmarking and mitigation of negative environmental impacts (OECD, 2022[72]).

Attempts to estimate the environmental costs of AI LMs are beginning

In spite of a lack of consensus among researchers on the amount of carbon emitted by AI language models, a number of experts and publications that attempt to estimate these. One expert estimated that integrating LLMs into search engine processes could require up to five times more computing power per search, with a commensurate increase in energy use and carbon emissions (Stokel-Walker, 2023[100]). Strubell et al. measured a big Transformer model's LM training and development costs in estimated CO2 emissions (Strubell, Ganesh and McCallum, 2019[101]). According to their research, training a large Transformer model produces 284 000 kilogrammes of CO2, which is equivalent to the CO2 emissions of five cars over their lifetime. However, Patterson et al. argued that the estimate was in fact 88 times lower (Patterson et al., 2021[102]). Patterson's study also found that training GPT-3 required the equivalent of over 552 000 kilogrammes of CO2 (Patterson et al., 2021[102]). Despite the significant scale of this statistic, they point out that LLMs such as GPT-3 bring potential energy benefits, one of which includes few-shot generalisation. Few-shot generalisation means that LLMs do not require re-training for every new task, in contrast to smaller models (Patterson et al., 2021[102]; Wang et al., 2020[103]).

Considering "model, machine, mechanisation and map" can reduce energy requirements

Few estimates measuring the energy footprint of AI systems exist. These estimates also rarely differentiate between AI training and use, or "inference", workloads. According to a study by Google in 2022, the energy use for machine learning workloads consistently represented less than 15% of its data centres' total energy use over the three-year period from 2019 to 2021 (Patterson, 2022[104]). Google found that machine learning energy requirements could be reduced by considering four aspects ("4M"): model, machine, mechanisation and map, as follows.

- Model: the computer software or algorithm to address a specific problem
- Machine: the computer hardware running the AI model
- Mechanisation: the data centre housing the machine
- Map: the geographic location of the data centre

Google released its Generalist Language Model (GLaM) in 2021, which reportedly optimised the 4Ms — improved models, ML-specific hardware, efficient data centres and data centres located in optimal geographies — approach to reduce the carbon emission of machine learning models by a factor of fourteen (Patterson, 2022[104]). Google released GLaM, which is reported to optimise the 4Ms — improved models, ML-specific hardware and efficient data centres — approach to reduce the carbon emission of machine learning models by a factor of fourteen (Patterson, 2022[104]).

Financing and other barriers to producers and users

The financial costs of training and using AI LLMs are substantial

The financial cost of substantial amounts of computing power and other compute resources are another barrier to training and using LLMs. For example, a 2020 study reported that OpenAI's GPT-3 model training would cost USD 4.6 million (Dickson, 2020[105]). According to a paper by AI21 Labs, it can cost between USD 2 500 and USD 50 000 to train a 110-million-parameter model, between USD 10 000 and USD 200 000 to train a 340-million-parameter model and between USD 80 000 and USD 1.6 million to train a 1.5-billion-parameter model (Sharir, Peleg and Shoham, 2020[106]). To put this into context, AI language models such as BLOOM and GPT-3 contain over 170 billion parameters.

Costs of AI LLMs are a barrier for SMEs, which can nonetheless fine-tune existing LMs

The high financial cost to train AI language models creates a significant barrier for smaller companies wishing to develop their own models in-house, while providing larger technology companies with a first mover advantage. At the same time, however, it is worthwhile to note that small and medium-sized businesses that may not have the resources to build their own model can nevertheless benefit from fine-tuning existing language models where appropriate (Wiggers, 2022[45]).

Costs of AI LLMs are also a barrier for models in minority languages

The cost barriers for training AI language models also apply to minority languages. Most of the leading language models are trained in English and developed by large technology companies. AI language models for minority languages are still few and far behind in terms of performance. There has nonetheless been progress. For example, Nagoudi et al. have developed IndT5, the first Transformer language model for Indigenous languages that is built on a corpus of ten indigenous languages and Spanish (Nagoudi et al., 2021[108]). Room for progress remains in terms of minority-language AI language models and thus a window of opportunity for companies seeking to penetrate untapped segments of the NLP market.

Human-centred values and fairness (Principle 1.2)

Considerations related to respect for human rights, democratic values and fairness, including democratic institutions, privacy, and human agency, are particularly important given the increasing capacity and reach of AI language models. The selection and curation of training data play a major role in this regard.

Misinformation and disinformation

The combination of AI LMs and disinformation can lead to wide scale deception

As language models are deployed across sectors, a major concern is the increasing use of AI language models to produce "disinformation" – deliberately fabricated untrue content designed to deceive, e.g. writing untrue texts and articles – or "misinformation" – false or misleading information that does not intend to harm, e.g. creating falsehoods for entertainment – that can damage public trust in democratic institutions (Lesher, Pawelec and Desai, 2022[109]).

Users may not be able to distinguish between human and AI-generated text

The rapid development and widespread availability of increasingly sophisticated and easy-to-use AI language models has dramatically increased the ease of producing mis- and disinformation and of manipulating opinions at scale, particularly with language models that behave "human-like". Language models and other generative AI tools can enable the production of fake news, deepfakes, and other forms of manipulated content that may be impossible to distinguish from real ones, yet can threaten democracy, social cohesion, and public trust in institutions. The combination of AI and disinformation can lead to deception at a wide scale that traditional approaches like fact checking, user education and media literacy, or detection tools would be challenged to address.

Yet LMs can also help detect some "untruths" online

AI language models have also in the past been able to help detect "untruths" online. Efforts in this regard include European Union's (EU) Fandango project, where researchers built software tools using NLP to help journalists and fact-checkers detect and fight fake news and disinformation based on semantic features that are characteristic of fake news (European Commission, 2018[110]). Another EU-financed project is GoodNews, which aims to build the technological capability for algorithmic fake news detection in social media by analysing the patterns of news spread in social networks and the observation that fake news is shared online differently from real news stories (for exmaple, with far more "shares" than "likes" on Facebook) (European Commission, 2020[111]). Based on these patterns, GoodNews attaches a credibility score to news items.

New viable solutions to address generative AI's impact on mis and disinformation are crucial

Without appropriate guardrails for LMs and other types of generative AI, they can exacerbate mis- and dis-information and reinforce harmful stereotypes. The OECD and other intergovernmental organisations participating in the globalpolicy.ai coalition have been discussing the possibility of launching a Global Challenge on Digital Trust with several other partners to deepen understanding of the risk that generative AI poses for mis and disinformation. The goal of the challenge would be to incentivise stakeholders around the world to develop innovative and viable solutions to mitigate the risks of dis and misinformation for generative AI.

Fairness and bias

AI language models can replicate biases present in their training data

AI language models can replicate stereotypes and discrimination through biases contained in language resources used as training data and in language models themselves (including pre-trained models). Often, language resources on which AI language models are trained reflect data about dominant social groups, which can raise diversity and inclusion concerns. Using balanced training datasets that represent different groups more equally is essential to minimise biases of AI language models. Model validation and verification is also crucial to assess and eliminate biases before a system's deployment.

Biases in language models include historical bias (e.g. pre-existing patterns in the training data and societal bias); representation bias (e.g. incomplete information due to missing variables, inadequate sample size, etc.); measurement bias (e.g. omission or inclusion of variables that should or should not be in the model); methodological and evaluation bias (e.g. errors in the use of evaluation metrics); monitoring bias (e.g. inappropriate interpretation of a system's outputs during monitoring); and feedback loops (e.g. a few popular items are recommended frequently to users, creating a feedback loop) (OECD, 2023[112]).

Larger parameter count can decrease certain types of bias but cause other issues

The number of parameters used to train a model can also lead to biases. In general, the greater the number of parameters, the less the model is exposed to certain types of bias (omitted-variable bias) (Goel, 2020[113]). However, a greater number of parameters often increases a model's privacy considerations and energy requirements.

Tools are being developed to help address biased or toxic data and feature on oecd.ai/tools

Approaches to auditing AI language models include data representativeness audits, using bias metrics such as equalised odds, and human evaluators to determine whether the results generate any type of bias or discrimination (Holistic AI, 2023[114]). For example, OpenAI uses machines and humans to monitor content used and produced by ChatGPT (Hsu and Thompson, 2023[115]). Human AI trainers and feedback from users identify and filter out toxic training data and teach ChatGPT to produce better responses.

Some researchers are also using one AI language model to try to identify biases in another. For example, DeepMind published a paper on "red-teaming" language models, that is, resolving biases in a target language models by using test cases from another model (Perez et al., 2022[116]). This approach has limitations, as the underlying bias of the "red-teaming" model may conceal the real amount of bias and discrimination of the red-teamed model.

Privacy concerns

LMs can lead to privacy breaches through leaking or inference of private information

Privacy is another important consideration for LMs, particularly for LLMs. Given that AI language models process vast amounts of unstructured data, there is a concern that even normal use of data by a LM can lead to personal data breaches through inadvertent leaking or inference of private information (Weidinger et al., 2021[98]). Privacy issues can also be caused by security breaches. The potential theft of training data that can include personal data is yet another risk. AI language models can also facilitate both legitimate and illegitimate surveillance and censorship.

Developers of AI language models need training and tools

In the selection of training data, specific datasets may contain confidential information that may be illegal to process under privacy laws such as the EU's General Data Protection Regulation harmonises regional privacy laws (European Union, 2016[117]). Yet developers of AI language models may not have the necessary training and expertise to understand and apply these regulations. Practical tools are needed to help guide them. One such tool is PLOT4AI, an open-source library with models and methods to protect against 86 threats to privacy. The OECD.AI Catalogue of Tools and Metrics for Trustworthy AI (oecd.ai/tools) provides a one-stop shop for AI actors to share tools to address risks to the AI Principles, including risks to privacy in the LM context. Privacy-enhancing technologies as well as privacy-by-design processes are important in the development and use of AI language models.

Transparency and explainability (Principle 1.3)

The opacity and complexity of LMs challenges transparency and explainability

Most AI LMs today rely on neural networks -- sophisticated statistical modelling techniques that are opaque and complex –, posing a central challenge to transparency and explainability of LMs. Even those who develop AI language models often do not understand how variables are combined to make predictions. Neural networks are often referred to as "black boxes" due to their complex, multi-variable probabilistic correlations that are most often difficult for the human mind to understand without additional tools (OECD, 2022[72]).

Meaningful information regarding the development and use of LMs is needed

Transparency associated with the use of language models includes disclosing when they are being used, establishing clear guidance on appropriate use, and warnings for misuse. Gaining access to meaningful information regarding the development and use of LMs is important to foster a general understanding of these technologies. Fairly recent releases of open-source models such as BigScience's BLOOM and Meta's OPT-175B have provided substantial public information about these language models.

People adversely affected by an LM's outcome should be able to challenge the outcome

Explainability means enabling people affected by the outcome of an AI system to understand how it was arrived at and entails providing easy-to-understand information that can enable those adversely affected by an AI system to challenge the outcome. Explainability is a rising issue in AI language models due to their complexity and opacity, particularly when it is necessary to challenge an outcome. In general, the more complex a model is, the harder it is to explain (OECD, 2022[39]).

Robustness, security and safety (Principle 1.4)

The 2019 OECD AI Principles state that "*AI systems should be robust, secure and safe throughout their entire life cycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk*." Yet increasingly powerful AI language models raise significant policy challenges related to ensuring that these transformative and increasingly powerful future models are safe, robust, secure and beneficial for individuals and organisations who are adopting them rapidly and widely.

Lack of human understanding of LM internal operation can lead to unpredictability and inability to constrain behaviour

A few companies with commercial imperatives and significant advance regarding language models are understandably competing to develop and deploy useful popular tools. Yet lack of understanding of their functioning leads to unpredictability and inability to constrain behaviour. Policy makers must encourage all actors, notably researchers, to develop rigorous quality control methodologies and standards for systems to meet at all stages of their lifecycle, appropriate to the application context. Mechanisms include investing in these areas of research and establishing minimum requirements in public procurement, in addition to standards and regulations.

Addressing AI language model safety challenges should be prioritised

Proactive and human-centred approaches can be helpful in addressing the safety challenges of AI language model. For example DeepMind recently released research on building safer dialogue agents powered by language models (DeepMind, 2022[118]). Sparrow is an information-seeking dialogue agent trained to be helpful, accurate, and harmless (e.g. not giving advice that could lead to physical safety or breaking the law). Sparrow uses reinforcement learning from human feedback to increase the usefulness of its answers and its behaviour is constrained by simple rules around possibly harmful advice.

Language models can be misused to develop malicious software code or create personalised scams and fraud

AI language models can generate security risks if they are misused (Weidinger et al., 2021[98]):

- To develop *malicious software code* for viruses and malware and cyberattacks in a number of different programming languages for cybercrime.
- To create *personalised* scams or fraud at scale, e.g. by drafting very realistic text or impersonation of specific individuals or groups for phishing attacks, fraud or disinformation.
- Intended or unintended *leaking or inference* of private or sensitive information can occur, with serious consequences for individual or group safety and security in both physical and digital environments.

Open-source AI language models are double-edged swords in terms of digital security

The development and use of open-source AI language models are a double-edged sword in terms of digital security. On the one hand, a broader community of developers and users can proactively identify errors, correct them, block hacking attempts and launch regular updates. Security by design can help in this regard to develop ensure information integrity, transparency and confidentiality in language models. On the other hand, malicious actors can also identify errors and exploit model vulnerabilities (Sagar, 2021[119]).

Building the evidence base can inform discussions on the safety of AI language models

The widespread deployment of AI language models is already producing new data on their capabilities, risks and incidents, which can help inform discussions of future AI safety risks with empirical evidence. In particular, the OECD is developing an AI incidents monitor to identify incidents reported in the press in real-time from reputable sources and start building the evidence base to inform discussions on the safety of AI language models and other AI technologies.

Accountability (Principle 1.5)

As the capacity and impact of AI language models increase, so does the importance of accountability when developing, deploying and operating these solutions. Debate is ongoing about who should be accountable for the proper functioning of AI language models, including general-purpose systems, and how liabilities should be determined.

AI actors should manage AI risks throughout the AI system lifecycle

Discussions since early 2020 in the OECD.AI network of experts have underlined that accountable actors should manage risks to the AI Principles throughout the lifecycle of their systems. This applies to actors involved in the development, deployment and use of NLP systems. Four steps can help to manage AI risks throughout the lifecycle: (1) Defining scope, context, actors and criteria; (2) Assessing risks at individual, aggregate and societal levels; (3) Treating risks in ways commensurate to cease, prevent or mitigate adverse impacts; and (4) Governing the risk management process (OECD, 2023[112]).

Useful resources exist but focused international and multi-stakeholder effort is needed on quality control and standards enabling accountability and control of powerful models

Several initiatives exist to develop industry standards, codes, and regulations to promote accountability in AI, including for AI language models. This includes existing OECD frameworks, such as the OECD AI Principles, the AI system lifecycle, and the OECD Framework for Classifying AI Systems. The proposed EU AI Act, currently under discussion, includes provisions pertaining to general-purpose AI systems.

Furthermore, some developers of AI language models, such as BLOOM, have created playbooks, model cards and data cards for their developer community, containing guidance on good practices. The recently-launched Catalogue of Tools & Metrics for Trustworthy AI (oecd.ai/tools) provides a one-stop shop for AI actors to share approaches, mechanisms and practices to implement trustworthy AI in a comparable and easily accessible manner, including tools to promote accountability for the development and use of large language models (OECD, 2023[120]).

Humans tend to trust AI system outputs, risking over-reliance

A well-known risk related to accountability is the human tendency to trust AI systems' recommendations somewhat blindly, including in critical decision-making. As AI language models become increasingly accurate, there is a risk of over relying on them despite their flaws. At the same time, several types of narrow AI systems may help improve humans' ability to operate or interpret AI systems and discourage overreliance by, for example, using AI systems specifically designed to help human operators interpret or critique certain answers provided by other AI systems.

Guardrails may be called for to control some forms of powerful language models

Looking to the future, guardrails may be called for to control some forms of powerful language models that can directly affect the real world. A question with significant societal implications is whether powerful language models should be able to take actions directly, such as sending emails, making purchases, and posting on social media, as opposed to their current use as passive question-answering systems.

Investing in R&D (Principle 2.1)

Some analysts estimate VC investment in generative AI increased by 500% in just 2 years

In 2022, as with the global VC market, overall VC investments in AI firms experienced a significant drop of over 40%. This reflects broader VC trends, with investors exercising more caution after the tech boom of the COVID-19 pandemic years, rising interest rates, and inflationary pressures. An exception to the cooling trend in VC investments is generative AI, which some analysts estimate has increased in VC investments by nearly 500% in just 2 years (from USD 230 million in 2020 to USD 1.37 billion in 2022) (Temkin, 2022[121]).

More benchmarks and assessment techniques are called for

Investment in the research and development of benchmarks and techniques for assessing language models and data is needed to evaluate their performance on criteria including accuracy, security, safety, explainability, fairness and bias on different tasks and at different degrees of difficulty. The General Language Understanding Evaluation (GLUE) benchmark, for example, assesses the technical performance of AI language models and measures their performance on three categories: single-sentence tasks, similarity and paraphrasing and inference tasks.

R&D in more energy-efficient mechanisms to train and query language models is critical

Given the environmental and financial computational costs associated with the development and training of NLP, R&D on more efficient mechanisms to train and query language models is also critical.

Co-operation and foresight are needed more than ever

AI language models are rapidly evolving and improving. Close cooperation on foresight and forecasting research to identify and anticipate future developments are needed to mitigate possible future risks to society. Progress in AI language models may happen at a steady pace but also through significant hardware and software breakthroughs that could affect the trajectory of AI progress in previously unexpected manners. Societal implications require major concertation and cooperation between all stakeholders, and particularly between researchers and policy makers.

Fostering a digital ecosystem (Principle 2.2)

Access to compute capacity is currently a pre-requisite to develop large language models

The development and deployment of AI language models and applications often requires extraordinary amounts of data and computing power. However, these resources come at a significant cost and are not accessible to everyone. Based on current trends, AI language models can be expected to continue to use greater and greater amounts of input data, parameters and compute, to perform increasingly difficult tasks,. Yet if this trend persists, a few large companies could continue to dominate the market as smaller companies or academic institutions may not have the resources to create or train large enough language models to compete.

"Ground truth" data, often human-based and contextual, is essential to language model training

Current AI language models often rely on human-labelled data as the "ground truth", which limits scalability. To overcome this challenge, developing ways to produce ground truth data at scale is vital. Some are attempting to teach AI systems "unaligned physics." For example, by understanding momentum and impact, an AI model could understand whether and why someone was injured in a car crash – instead of relying on human labelling. Yet, given societal and cultural differences, "ground truth" for AI language models may often be contextual. The use of synthetic data has also gained widespread adoption in the ML community, as it allows to simulate scenarios that are difficult to observe or replicate in real life and improve representativeness of incomplete data sets.

Yet language models decreasing overall quality of data online is a risk that needs to be mitigated

A data-related risk posed by AI language models is the possibility of producing low-quality or inaccurate outputs that would then decrease the overall quality of data online. For example, low-quality AI-based translations may lower the aggregate quality of online translations and, in consequence, online content. Training models on synthetic text is also typically less accurate than training on human text. It is thus possible that the data produced by AI language models could harm the online "commons" and produce a vicious cycle whereby AI systems are trained on ever lower quality data produced by AI language models themselves.

Language resources development in specialised domains and minority languages is promising

Investing in language resources and data repositories in minority languages is important, as most existing open source language models and datasets cover the more commonly used languages – namely, English, German, Spanish and French. Many national governments are already engaging in concrete initiatives to this end (see National AI policies and initiatives for language models).

Opening access to language models, including via open-source licensing, is promising but risky

Opening access to AI language models can increase inclusiveness and participation in the AI ecosystem, allowing individuals, organisations, and communities to develop and use language models regardless of their financial resources. Open sourcing the underlying code of language models so that it is public and modifiable can help improve transparency and inclusiveness. However, open-source models come with risks, as they can be reused and manipulated by malignant actors.

Organisations and SMEs with less resources can fine-tune existing language models

Organisations and SMEs with fewer economic and technological resources can investigate fine-tuning existing language models to their operations, as well as improve customer service, e.g. with personalisation and interactive chatbots. A key obstacle to LM diffusion in firms is the availability of digitalised data and AI skills, but SMEs' use of AI could in some cases take place via specific AI modules of business operation and customer relation management (CRM) software like SAP, such that SMEs using AI may even be able to do so seamlessly, without even realising they are using AI.

Fostering an enabling policy environment (Principle 2.3)

AI language models should be included in national AI strategies and action plans

Including language models in national AI strategies and action plans is important to encourage their trustworthy development and deployment. Some governments have already started to engage in this effort. For example, in the UK, in Wales the government has unveiled the Welsh Language Technology Action Plan (OECD.AI, 2021[26]), while the Government of Spain has introduced its Plan for the Advancement of Language Technology (OECD.AI, 2021[24]).

Policy makers are considering different mechanisms to ensure safe and beneficial LMs

Reaping the benefits of language models and generative AI while addressing their challenges and risks requires an enabling policy environment and thoughtful mechanisms to ensure safe and beneficial LMs. Discussions are ongoing to identify seek the best ways to account for generative AI in hard and soft law. Some jurisdictions are taking a cross-sectoral “horizontal” approach to AI regulation (Canada, European Union), while others consider a more sectoral or “vertical” approach (United States, United Kingdom). Questions remain related to interoperability, defining high-risk AI systems, how to treat general-purpose AI, and effective enforcement mechanisms, among others.

AI-specific legislation is being considered in jurisdictions such as the proposed AI Act in Europe and the AI and Data Act in Canada. The proposed AI Act would include obligations for providers of General Purpose AI systems that would be considered high-risk. Others are adapting existing legislation and complementing it with policy guidance and frameworks for AI language models. Several experts have called for a pause in the development of more powerful AI language models than GPT-4 (Future of Life Institute, 2023[122]). Intergovernmental organisations (IGOs) also play a crucial role in helping policy makers share good practices and develop interoperable solutions.

Building human capacity and preparing for labour transitions (Principle 2.4)

AI LMs can automate more and more tasks, including traditionally highly skilled tasks

AI technologies are impacting a wide range of tasks and workers (Milanez, 2023[123]). AI language models can automate tasks such as transcription and translation, which may impact several occupations. Last-generation language models are increasingly being used to conduct more complex and creative tasks, like developing and debugging software code and writing music or poetry. Increasingly capable AI systems, such as AI language models and agent-like AI systems, are raising concerns about job displacement, including of high-wage workers such as programmers and ensuing economic and social disruption. By way of illustration, recent advances in AI language models to automate code writing have been impressive. Examples of solutions include DeepCoder (developed by Microsoft and Cambridge University), AlphaCode (developed by DeepMind), Codex (OpenAI's natural-language coding project), Co-pilot (developed by GitHub) and ChatGPT (developed by OpenAI).

Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society

Governments should empower people to effectively use and interact with AI language models across the breadth of applications. This includes equipping them with the necessary skills and enabling social dialogue and lifelong learning programmes to ensure a fair transition for workers affected by the development and use of AI language models.

In addition, some AI language models leverage human inputs to clean and label training data or otherwise contribute human judgment to AI systems. Workers, oftentimes low-skilled, underpaid and hired under precarious conditions (i.e. "ghost workers") are tasked with enriching the system's training data and improving the quality of the system's outcomes. Due to the complexity of the supply chain, even well-meaning firms, consumers and end-users may be unaware of the degree of "ghost work" or other human involvement in developing AI language models. Worker consultations may offer advantages when it comes to ensuring that job quality is maintained and even enhanced by AI and that the gains of AI language models are shared with all workers (Lane and Williams, 2023[124]).

International, interdisciplinary, and multi-stakeholder co-operation (Principle 2.5)

AI LMs can benefit from international, multi-disciplinary and multi-stakeholder cooperation

Research into trustworthy and beneficial AI language models is a collaborative effort that requires experts from different disciplines, including linguistics, computer science, cognitive psychology and policy makers. International research partnerships in this field also allow the share of knowledge and resources. International, interdisciplinary, and multi-stakeholder co-operation is required to prevent or mitigate harmful uses of AI language models, including the manipulation of opinions and the creation of mis- and disinformation at scale. Stakeholders, including policy makers, are already exploring the societal risks of AI language models, but much more work remains to develop solutions that can effectively mitigate their risks while fostering their beneficial development and adoption.

Regional and international fora can facilitate such co-operation

Collaboration can take the form of sharing best practices and lessons learned in regional or international fora. It can also take the form of joint initiatives in multilingual data and models. Furthermore, developing standards for AI language models or including AI language models in existing standards is important for interoperability between different systems and for the development and deployment of applications that work in a trustworthy manner across borders. The OECD and other international fora can leverage their convening power to shape standards that address risks stemming from the development and use of AI language models and generative AI more broadly.

References

AI4Bahrat (2022), *“Nilekani Center at AI4Bharat” was officially launched on 28th July at IIT Madras*. [33]

Amazon (2022), *20B-parameter Alexa model sets new marks in few-shot learning*, https://www.amazon.science/blog/20b-parameter-alexa-model-sets-new-marks-in-few-shot-learning. [59]

Anthropic (2023), *Introducing Claude*, https://www.anthropic.com/index/introducing-claude. [63]

Arabic Language Technology Group (2022), *Arabic Language Technology Group*, https://alt.qcri.org/. [34]

Barba, P. (2020), *Challenges in Developing Multilingual Language Models in Natural Language Processing (NLP)*, https://towardsdatascience.com/challenges-in-developing-multilingual-language-models-in-natural-language-processing-nlp-f3b2bed64739. [84]

Benaich, N. et al. (2022), *State of AI Report*, https://www.stateof.ai/ (accessed on 28 February 2023). [66]

Bilogur, A. (2020), *Transformer architecture, self-attention*, https://www.kaggle.com/code/residentmario/transformer-architecture-self-attention/notebook. [49]

Blanc., C. et al. (2022), *FlauBERT vs. CamemBERT: Understanding patient’s answers by a French medical chatbot*, https://www.sciencedirect.com/science/article/pii/S093336572200029X?via%3Dihub. [87]

Borgeaud, S. et al. (2022), *Improving language models by retrieving from trillions of tokens*, arXiv, https://arxiv.org/abs/2112.04426. [77]

Borgonovi, F., J. Hervé and H. Seitz (2023), *Not lost in translation: Machine translation technologies and their implications for the skills and labour market opportunities of language professionals and for broader society*, https://www.oecd.org/publications/not-lost-in-translation-e1d1d170-en.htm. [99]

Bovet, M. (1976), *Piaget’s Theory of Cognitive Development and Individual Differences*, https://link.springer.com/chapter/10.1007/978-3-642-46323-5_20. [127]

Bubeck, S. et al. (2023), “Sparks of Artificial General Intelligence: Early experiments with GPT-4”, *arXiv*, https://arxiv.org/pdf/2303.12712.pdf. [67]

Cabinet Office of Japan (2021), *総務省説明資料*, https://www8.cao.go.jp/cstp/ai/shin_ai/2kai/siryo1.pdf. [12]

Chen, X. et al. (2019), “Multi-Source Cross-Lingual Model Transfer: Learning What to Share”, https://doi.org/10.18653/v1/P19-1299. [92]

Chomsky, N. (1959), *Review of Skinner's Verbal Behavior*, http://www.ugr.es/~fmanjon/A%20Review%20of%20B%20%20F%20%20Skinner's%20Verbal%20Behavior%20by%20Noam%20Chomsky.pdf. [126]

CLARIN (2022), *The research infrastructure for language as social and cultural data*, https://www.clarin.eu/. [21]

CLARINO (2020), *Welcome to CLARINO Bergen Centre*, https://repo.clarino.uib.no/xmlui/. [20]

DeepMind (2022), *Building safer dialogue agents*, https://www.deepmind.com/blog/building-safer-dialogue-agents. [118]

Devlin, J. and M. Chang (2018), *Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing*, https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html. [52]

Dharmaraj, S. (2020), *Vietnam launches AI-based language applications*, https://opengovasia.com/vietnam-launches-ai-based-language-applications/. [38]

Dickson, B. (2020), *The GPT-3 economy*, https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/. [105]

Digwatch (2021), *The government of India announces the National Language Translation Mission to enhance access to digital content*, https://dig.watch/updates/government-india-announces-national-language-translation-mission-enhance-access-digital. [32]

Dilmegani, C. (2022), *In-Depth Guide into Natural Language Understanding in 2022*, https://research.aimultiple.com/nlu/. [41]

Estonian Ministry of Education and Research (2018), *The Language Technology Research and Development Program "Estonian Language Technology 2018-2027" of the Ministry of Education and Research*, https://www.hm.ee/sites/default/files/estonian_language_tech. [4]

European Commission (2020), *Fake news detection in social networks using geometric deep learning*, https://cordis.europa.eu/project/id/812672. [111]

European Commission (2018), *The Fandango project: Fake news discovery and propagation from big data analysis and artificial intelligence Operations*, https://cordis.europa.eu/project/id/780355 (accessed on 5 April 2023). [110]

European Language Equality (2023), *European Language Equality: Developing an agenda and a roadmap for achieving full digital language equality in Europe by 2030*, https://european-language-equality.eu/about/ (accessed on 5 April 2023). [44]

European Language Equality (2021), *D3.1: Report on existing strategic documents and projects in LT/AI*, https://european-language-equality.eu/wp-content/uploads/2021/05/ELE___Deliverable_D3_1.pdf. [85]

European Language Grid (2023), *European Language Grid (release 3): Towards the Primary Platform for Language Technologies in Europe*, https://live.european-language-grid.eu/ (accessed on 5 April 2023). [43]

European Union (2016), *Regulation (EU) 2016/679 of the European Parliament and of the Council*, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. [117]

Future of Life Institute (2023), *Pause Giant AI Experiments: An Open Letter*, https://futureoflife.org/open-letter/pause-giant-ai-experiments/. [122]

Gaidienė, A. and A. Tamulionienė (2022), *Report on the Lithuanian Language*, https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_23__Language_Report_Lithuanian_.pdf. [15]

Goel, M. (2020), *The Bias-Variance Trade-Off : A Mathematical View*, https://medium.com/snu-ai/the-bias-variance-trade-off-a-mathematical-view-14ff9dfe5a3c. [113]

Google (2021), *More Efficient In-Context Learning with GLaM*. [53]

Government of India (2021), *Technology Development for Indian Languages (TDIL)*, https://www.meity.gov.in/content/technology-development-indian-languages-tdil. [30]

Government of South Africa (2019), *Government Establishes A New Digital Centre To Promote Indigenous Languages*, https://www.dst.gov.za/index.php/media-room/latest-news/2885-government-establishes-a-new-digital-centre-to-promote-indigenous-lang. [35]

Gupta, K. (2022), *Amazon's 20B-Parameter Alexa Model Sets New Marks In Few-Shot Learning Along With Low Carbon Footprint During Training (One-Fifth of GPT-3's)*, https://www.marktechpost.com/2022/08/03/amazons-20b-parameter-alexa-model-sets-new-marks-in-few-shot-learning-along-with-low-carbon-footprint-during-training-one-fifth-of-gpt-3s/. [60]

Hoffman, J. et al. (2022), *Training Compute-Optimal Large Language Models*, https://arxiv.org/abs/2203.15556. [74]

Holistic AI (2023), *The Rise of Large Language Models: Galactica, ChatGPT, and Bard*, https://www.holisticai.com/blog/language-models-galactica-chatgpt-bard. [114]

Hsu, T. and S. Thompson (2023), "Disinformation Researchers Raise Alarms About A.I. Chatbots", https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html. [115]

Hugging Face (2023), *ALBERT*, https://huggingface.co/docs/transformers/model_doc/albert (accessed on 5 April 2023). [80]

Hugging Face (2023), *Bart*, https://huggingface.co/transformers/v3.0.2/model_doc/bart.html#:~:text=Bart%20uses%20a%20standard%20seq2seq,right%20decoder%20(like%20GPT) (accessed on 5 April 2023). [79]

Hugging Face (2023), *Bigscience T0*, https://huggingface.co/bigscience/T0 (accessed on 5 April 2023). [81]

Hugging Face (2022), *Illustrating Reinforcement Learning from Human Feedback (RLHF)*, https://huggingface.co/blog/rlhf. [50]

Hugging Face (2022), *Introducing The World's Largest Open Multilingual Language Model: BLOOM*, https://bigscience.huggingface.co/blog/bloom. [61]

Hugging Face (2021), *BERTIN Project*, https://huggingface.co/bertin-project. [89]

IBM (2020), *Speech Recognition*, https://www.ibm.com/cloud/learn/speech-recognition. [96]

iFlytek (2020), *iFLYTEK signs Arabic-Chinese Language Translation Research Agreement with Egyptian Ministry of Communications and Information Technology*, [29] https://www.prnewswire.com/ae/news-releases/a-i-connects-the-silk-road-iflytek-signs-arabic-chinese.

Kaplan, J. et al. (2020), *Scaling Laws for Neural Language Models*, https://arxiv.org/abs/2001.08361. [73]

Koradiya, R. (2019), *Differentiate Between NLP, NLG, and NLU*, https://iconflux.com/blog/differentiate-between-nlp-nlg-and-nlu. [93]

Kückens, J. (2022), *Spanish government invests 1.1 billion Euros into "New Language Economy", including co-official languages*, https://european-language-equality.eu/2022/03/04/spanish-government-invests-into-new-language-economy/. [25]

Kummervold, P. et al. (2021), "Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model", *NoDaLiDa*, https://doi.org/10.48550/arXiv.2104.09617. [90]

Kutuzov, A. et al. (2021), "Large-Scale Contextualised Language Modelling for Norwegian", *NoDaLiDa*, https://aclanthology.org/2021.nodalida-main.4. [91]

Lane, M. and M. Williams (2023), "Defining and classifying AI in the workplace", *OECD Social, Employment and Migration Working Papers*, No. 290, OECD Publishing, Paris, https://doi.org/10.1787/59e89d7f-en. [124]

Lee, S. (2021), *Kakao Brain unveils 'KoGPT', a Korean-specific AI language model*, Smart Times, https://www.smarttimes.co.kr/news/articleView.html?idxno=1371 (accessed on 5 April 2023). [78]

Lesher, M., H. Pawelec and A. Desai (2022), *Disentangling untruths online: Creators, spreaders and how to stop them*, https://goingdigital.oecd.org/data/notes/No23_ToolkitNote_UntruthsOnline.pdf. [97]

Lesher, M., H. Pawelec and A. Desai (2022), "Disentangling untruths online: Creators, spreaders and how to stop them", *OECD Going Digital Toolkit Notes*, No. 23, OECD Publishing, Paris, https://doi.org/10.1787/84b62df1-en. [109]

LT-Innovate (2023), *Finnish Government Termbank Valter*, http://www.lt-innovate.org/lt-observe/resources/finnish-government-termbank-valter. [9]

LT-Innovate (2023), *GERMANY*, http://www.lt-innovate.org/lt-observe/germany. [10]

LT-Innovate (2023), *HUNGARY*, http://www.lt-innovate.org/lt-observe/hungary. [11]

LT-Innovate (2023), *SLOVENIA*, http://www.lt-innovate.org/lt-observe/slovenia. [23]

Marconi, G. et al. (2020), "What matters for language learning? The questionnaire framework for the PISA 2025 Foreign Language Assessment", *OECD Education Working Papers*, No. 234, OECD Publishing, Paris, https://doi.org/10.1787/5e06e820-en. [40]

Marking, M. (2020), *Egypt Inks Machine Translation, NLP Deal with China's iFlytek*, https://slator.com/egypt-inks-machine-translation-nlp-deal-with-chinas-iflytek/. [28]

Martin, L. et al. (2020), *CamemBERT: a Tasty French Language Model*, https://aclanthology.org/2020.acl-main.645.pdf. [86]

Masakhane (2020), *A grassroots NLP community for Africa, by Africans*. [36]

Meta (2023), *Introducing LLaMA: A foundational, 65-billion-parameter large language model*, [56] https://ai.facebook.com/blog/large-language-model-llama-meta-ai/.

Meta (2022), *Democratizing access to large-scale language models with OPT-175B*, https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/. [54]

Milanez, A. (2023), "The impact of AI on the workplace: Evidence from OECD case studies of AI implementation"*, OECD Social, Employment and Migration Working Papers*, No. 289, OECD Publishing, Paris, https://doi.org/10.1787/2247ce58-en. [123]

Ministère de la Culture (2021), *AI and language technology in France*, https://lr-coordination.eu/sites/default/files/LRB/LRB-11/2.2%20-%20Pr%C3%A9sentation%2011th%20LRB%20IA%20in%20France%20v2_EN.pdf?lang=sv. [6]

Ministry of Foreign Affairs of Denmark (2021), *Denmark to stregtnen opportunities for NLP businesses*, https://investindk.com/insights/denmark-to-strenghten-opportunities-for-nlp-businesses. [2]

Ministry of Internal Affairs and Communications of Japan (2021), *Further Upgrading of Multilingual Translation System Supporting the Labor Field*, https://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/pressrelease/2021/3/30_04.html. [13]

Moberg, J. (2020), *A deep dive into multilingual NLP models*, https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models. [82]

Multilingual (2021), *Estonian Government to Develop Central Translation Platform*, https://multilingual.com/estonia-translation-platform/. [5]

Nagoudi, E. et al. (2021), *IndT5: A Text-to-Text Transformer for 10 Indigenous Languages*, https://arxiv.org/abs/2104.07483. [108]

National Library of Norway (2023), *The Norwegian Language Bank*, https://www.nb.no/sprakbanken/en/sprakbanken/. [18]

National Research Council Canada (2021), *Canadian Indigenous languages technology project*, https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project. [1]

Nordic Language Processing Laboratory (2021), *Nordic Language Processing Laboratory*, http://wiki.nlpl.eu/index.php/Home. [22]

OECD (2023), "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI"*, OECD Digital Economy Papers*, No. 349, OECD Publishing, Paris, https://doi.org/10.1787/2448f04b-en. [112]

OECD (2023), *Catalogue of Tools & Metrics for Trustworthy AI*, https://oecd.ai/en/catalogue/ (accessed on 5 April 2023). [120]

OECD (2022), "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint"*, OECD Digital Economy Papers*, No. 341, OECD Publishing, Paris, https://doi.org/10.1787/7babf571-en. [72]

OECD (2022), "OECD Framework for the Classification of AI systems"*, OECD Digital Economy Papers*, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. [39]

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, https://doi.org/10.1787/eedfee77-en. [70]

OECD.AI (2021), *Spain: National plan for the advancement of LTS*, https://oecd.ai/en/dashboards/policy-initiatives/http:%2F%2Faipo.oecd.org%2F2021-data-policyInitiatives-16665 (accessed on 5 April 2023). [24]

OECD.AI (2021), *The United Kingdom: Welsh Language Technology Action Plan*, https://oecd.ai/en/dashboards/policy-initiatives/http:%2F%2Faipo.oecd.org%2F2021-data-policyInitiatives-26880 (accessed on 5 April 2023). [26]

OpenAI (2023), *GPT-4*, https://openai.com/research/gpt-4. [51]

Pandey, V. (2022), *Power of Natural Language Processing (NLP) and its Applications in Business*, https://www.linkedin.com/pulse/power-natural-language-processing-nlp-its-business-dr-vivek-. [95]

Patterson, D. (2022), *Reducing the carbon emissions of AI, OECD.AI Wonk Blog*, https://oecd.ai/en/wonk/reducing-ai-carbon-emissions (accessed on 15 March 2023). [104]

Patterson, D. et al. (2021), *Carbon Emissions and Large Neural Network Training*, https://arxiv.org/pdf/2104.10350.pdf. [102]

Paullada, A. (2021), *Data and its (dis)contents: A survey of dataset development and use in machine learning research*, https://www.sciencedirect.com/science/article/pii/S2666389921001847. [71]

Perez, E. et al. (2022), *Red Teaming LMs with LMs*, DeepMind, https://arxiv.org/pdf/2202.03286.pdf. [116]

Piaf (2023), *Piaf: Pour des IA francophones (beta)*, https://piaf.etalab.studio/. [7]

Pires, T., E. Schlinger and D. Garrette (2019), *How multilingual is Multilingual BERT?*, https://research.google/pubs/pub48247/. [83]

Press Trust of India (2021), *Govt working on real-time translation tool for Indian languages: MeitY*, https://www.theweek.in/news/india/2021/06/19/govt-working-on-real-time-translation-tool-for-indian-languages-meity.html. [31]

Raj, D. (2017), *Trends in Semantic Parsing — Part 1*, https://medium.com/explorations-in-language-and-learning/trends-in-semantic-parsing-part-1-ba11888523cb. [94]

Rodrigo et al. (2021), *Report on the state of the art in LT and Language-centric AI*. [42]

Rodriguez, J. (2021), *Five Key Facts Wu Dao 2.0: The Largest Transformer Model Ever Built*, https://medium.com/dataseries/five-key-facts-wu-dao-2-0-the-largest-transformer-model-ever-built-19316159796b. [58]

Rojas, J. et al. (2020), *BETO: Spanish BERT*, https://github.com/dccuchile/beto. [88]

Romero, A. (2021), *Wu Dao 2.0: A Monster of 1.75 Trillion Parameters*, https://towardsdatascience.com/gpt-3-scared-you-meet-wu-dao-2-0-a-monster-of-1-75-trillion-parameters-832cd83db484. [27]

Rowe, M. and A. Weisleder (2020), *Language Development in Context*, https://www.researchgate.net/publication/344404419_Language_Development_in_Context. [128]

Russell, S. and P. Norvig (2016), *Artificial Intelligence: A Modern Approach*, Pearson Education, Inc. [69]

Sagar, R. (2021), *Security Risks Of Open Source Software*, https://analyticsindiamag.com/security-risks-of-open-source-software/. [119]

Schick, T. and H. Schütze (2021), *It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners*, https://arxiv.org/abs/2009.07118. [75]

SCRIBE (2021), *SCRIBE – Machine transcription of Norwegian conversational speech*, https://scribe-project.github.io/. [19]

Sevilla, J. et al. (2022), "Compute Trends Across Three Eras of Machine Learning", https://arxiv.org/abs/2202.05924 (accessed on 22 February 2023). [65]

Sharir, O., B. Peleg and Y. Shoham (2020), *The Cost of Training NLP Models: A Concise Overview*, https://arxiv.org/pdf/2004.08900.pdf. [106]

Simon, J. (2021), *Large Language Models: A New Moore's Law?*, https://huggingface.co/blog/large-language-models. [64]

Simoulin, A. and B. Crabbé (2021), *How Many Layers and Why? An Analysis of the Model Depth in Transformers*, HAL Open Science, https://hal.archives-ouvertes.fr/hal-03601412/document. [46]

Skinner, B. (1957), *Verbal behavior*, Appleton Century, https://doi.org/10.1037/11256-000. [125]

Stokel-Walker, C. (2023), *The Generative AI Race Has a Dirty Secret*, Wired, https://www.wired.co.uk/article/the-generative-ai-search-race-has-a-dirty-secret. [100]

Strubell, E., A. Ganesh and A. McCallum (2019), *Energy and Policy Considerations for Deep Learning in NLP*, https://arxiv.org/pdf/1906.02243.pdf. [101]

Sun, Y. et al. (2021), *ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation*, arXiv, https://arxiv.org/abs/2107.02137. [76]

Synced (2022), *Meta AI Open-Sources a 175B Parameter Language Model: GPT-3 Comparable Performance at One-Seventh the Compute Cost*, https://syncedreview.com/2022/05/06/meta-ai-open-sources-a-175b-parameter-language-model-gpt-3-comparable-performance-at-one-seventh-the-compute-cost/. [55]

Synced (2021), *Baidu's Knowledge-Enhanced ERNIE 3.0 Pretraining Framework Delivers SOTA NLP Results, Surpasses Human Performance on the SuperGLUE Benchmark*, https://medium.com/syncedreview/baidus-knowledge-enhanced-ernie-3-0-7eb37bf098dd. [57]

Tapsai, C., H. Unger and P. Meesad (2021), *Thai Word Segmentation*, Springer, https://doi.org/10.1007/978-3-030-56235-9_2. [37]

Temkin, M. (2022), *VCs try to parse through the 'noise' of generative AI*, https://pitchbook.com/news/articles/generative-ai-venture-capital-investment. [121]

Tilde (2022), *Tilde Wins a Major Contract to Provide Machine Translation Services to the Finnish Government*, https://slator.com/tilde-wins-a-major-contract-to-provide-machine-translation-services-to-the-finnish-government/. [8]

Tilde (2021), *Tilde Develops a Unique Public Machine Translation Platform*, [16] https://slator.com/tilde-develops-a-unique-public-machine-translation-platform/.

UNESCO (2020), *The Norwegian Language Bank ("Språkbanken") – National Infrastructure for Language Technology*, https://fr.unesco.org/creativity/policy-monitoring-platform/norwegian-language-bank. [17]

University of Copenhagen (2023), *Centre for Language Technology*, https://cst.ku.dk/english/. [3]

Vaswani, A. et al. (2017), *Attention Is All You Need*, https://arxiv.org/abs/1706.03762. [48]

Viksna, R. et al. (2022), *Assessing Multilinguality of Publicly Accessible Websites*, European Language Resources Association, https://aclanthology.org/2022.lrec-1.227. [14]

Wang, Y. et al. (2020), *Generalizing from a few examples: A survey on few-shot learning*, https://arxiv.org/abs/1904.05046. [103]

Weidinger, L. et al. (2021), *Ethical and social risks of harm from Language Models*, https://arxiv.org/abs/2112.04359. [98]

Wellman, H. (1992), *The child's theory of mind*, The MIT Press. [68]

Wiggers, K. (2022), *The emerging types of language models and why they matter*, https://techcrunch.com/2022/04/28/the-emerging-types-of-language-models-and-why-they-matter/?renderMode=ie11&guccounter=1. [45]

Yandex (2022), *Yandex publishes YaLM 100B, the largest GPT-like neural network in open source*, https://yandex.com/company/press_center/press_releases/2022/2022-23-06. [62]

Yang, Z. et al. (2022), *CINO: A Chinese Minority Pre-trained Language Model*, https://arxiv.org/abs/2202.13558. [107]

Ye, A. (2020), *Breaking Down the Innovative Deep Learning Behind Google Translate*, https://medium.com/analytics-vidhya/breaking-down-the-innovative-deep-learning-behind-google-translate-355889e104f1. [47]

Annex A. Language acquisition and learning

Language is the main vector of human communication. It enables people to learn from, interact, and connect with others. “Natural language” refers to any language that has evolved in humans through use and repetition without conscious planning or premeditation. Natural language processing (NLP) refers to computer programs and tools that automate natural language functions by analysing, producing, modifying, or responding to human texts and speech. NLP and human language learning have similarities in terms of skills used. Benchmarks are being developed to compare the performance of different language models as well as to compare them to human performance. Different theories of language acquisition and learning explain the complexities of human language learning that should be considered in developing and deploying NLP systems that most closely mimic human performance.

Over the last fifty years, several theories have been put forward to explain the process by which people at their early stages of life acquire a language:

- the *Behaviourist theory (Skinner)* (Skinner, 1957[125]) suggests that a child imitates the language of its parents or caregivers;
- the *Innateness theory (Chomsky)* (Chomsky, 1959[126]) puts forward that human brain structures naturally permit the capacity to learn and use languages, making language acquisition a biologically determined process that uses neural circuits in the brain, which have evolved to contain linguistic signals;
- the *Cognitive theory (Piaget)* (Bovet, 1976[127]) places acquisition of language within the context of a child's mental or cognitive development. In other words, a child has to understand a concept before s/he can acquire the particular language form, which expresses that concept; and
- the *Input or Interactionist theory (Bruner)* (Rowe and Weisleder, 2020[128]) emphasises the importance of the language input children acquire from their caregivers. In this context, language exists for the purpose of communication and can only be learned through interacting with other people.

Annex B. GPT-3 in the context of the OECD Framework for the Classification of AI Systems

People & Planet

Core characteristic	Survey question	Response
Users of AI system	What is the level of competency of users who interact with the system?	Amateur
Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?	Workers (e.g. could lead to automation of some tasks), consumers
Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?	Optional / can opt out
Human rights and democratic values	Can the system's outputs impact fundamental human rights?	*Possible impact on*: - rule of law, absence of arbitrary sentencing - freedom of thought, conscience and religion - equality and non-discrimination - quality of democratic institutions (e.g., free elections)
Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?	*Possible impact on*: - work and job quality - education
{Displacement potential}	*Could the system automate tasks that are or were being executed by humans?*	*TBD*

Economic Context

Core characteristic	Survey question	Response
Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?	Section J: Information and Communication (per ISIC REV 4)
Business function	What business function(s) or functional areas is the AI system employed in (e.g. sales, customer service, human resources)?	Any
Business model	Is the system a for-profit use, non-profit use, or public service system?	For-profit use – other model (e.g. business intelligence) or non-profit use (e.g. research, journalism)
Impacts critical functions / activities	Would the disruption of the system's function or activity affect essential services?	No
Breadth of deployment	Is the AI system deployment a pilot, narrow, broad, or widespread?	Narrow deployment
{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)?*	*System prototype demonstration in operational environment – TRL 7*

Data & Input

Core characteristic	Survey question	Response
Detection and collection	Are the data and input collected by humans, automated sensors, both?	Collected by humans and automated sensing devices (e.g. collected by humans with subsequent filtering by machines and humans)
Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?	Observed and derived
Dynamic nature	Are the data dynamic, static, dynamic updated from time to time or real-time?	Dynamic data updated from time to time
Rights associated with data and input	Are the data proprietary (privately held), public (no intellectual property rights) or personal data (related to identifiable individual)?	Public and proprietary
Identifiability of personal data	If personal data, are they anonymised, pseudonymised?	N/A
{Data quality and appropriateness }	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*	*Noisy data, that is, by design, highly representative and diverse with regard to a large part of (predominantly English) text and code found on the Internet; appropriate data*
{Structure of the data and input}	*Are the data structured, semi-structured, complex structured or unstructured?*	*Unstructured data*
{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*	*Non-standardised*
{Scale}	*What is the dataset's scale?*	*Very large*

AI Model

Core characteristic	Survey question	Response
Model information availability	Is any information available about the system's model?	Yes
AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	Statistical (data-driven)
{Rights associated with model}	*Is the model open-source or proprietary, self, or third-party managed?*	*Proprietary*
{Discriminative or generative}	*Is the model generative, discriminative or both?*	*Generative*
{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*	*One*
Model-building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning or through reinforcement learning?	Acquisition from data, augmented by human-encoded knowledge
Model evolution in the field ^{ML}	Does the model evolve and / or acquire abilities from interacting with data in the field?	Evolution during operation through passive interaction
Central or federated learning ^{ML}	Is the model trained centrally or in a number of local servers or edge devices?	Central
{Model development and maintenance}	*Is the model universal, customisable, or tailored to the AI actor's data?*	*Context-dependent*
{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*	*Deterministic*
Transparency and explainability	*Is information available to users to allow them to understand model outputs?*	*Context-dependent*

Task & Output

Core characteristic	Survey question	Response
Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?	Reasoning with knowledge structures, interaction support, recognition, personalisation
{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*	*Yes*
Action autonomy	How autonomous are the system’s actions and what role do humans play?	Low autonomy
Core application area(s)	Does the system belong to a core application area such as human LTs, computer vision, automation and / or optimisation or robotics?	Human LTs
{Evaluation methods}	*Are there standards or methods available for evaluating system output?*	*TBD*