EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

AID 2030 – Artificial Intelligence Data Kit 2030

Authors	Svetla Koeva, Valentina Stefanova (Appendices)
Organisation	Institute for Bulgarian Language Prof. Lyubomir Andreychin

Dissemination level	Public
Date	15-05-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	AID 2030 – Artificial Intelligence Data Kit 2030
Authors	Svetla Koeva, Valentina Stefanova (Appendices)
Organisation	Institute for Bulgarian Language Prof. Lyubomir Andreychin
Type	Report
Number of pages	54
Status and version	Final
Dissemination level	Public
Date of delivery	15-05-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial intelligence
AID2030	Artificial Intelligence Data Kit 2030
API	Application Programming Interface
BIG-bench	Beyond the Imitation Game Benchmark
BLARK	Basic LAnguage Resource Kit
BLOOM	BigScience Large Open-science Open-access Multilingual Language Model
BLUE	BiLingual Evaluation Understudy
CB	Commitment Bank
CLUE	Chinese Language Understanding Evaluation
COPA	Choice of Plausible Alternatives
CPM	Chinese pre-trained language model
DeepStruct	Deep Structural Prediction
ELDA	Evaluations and Language resources Distribution Agency
ELE	European Language Equality
ELE2	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019–2022)
ELRA	European Language Resources Association
ELSNET	European Network of Excellence in Language and Speech
FewCLUE	Chinese Few-shot Learning Evaluation
FLOPs	Floating point operations
GLaM/GLM	General Language Model
GLUE	General Language Understanding Evaluation
GPU	Graphics Processing Unit
GPT	Generative Pre-trained Transformer
HELM	Holistic Evaluation of Language Models
IPR	Intellectual property rights
LaMDA	Language Models for Dialog Applications
LR	Language resources
LLaMA	Large Language Model Meta AI
LLD	Large language dataset

LLM	Large language model
LT	Language Technology/Technologies
MIP	Multi-task Instruction Pre-training
MIT	Massachusetts Institute of Technology
MMLM	Multilingual Masked Language Modeling
MMLU	Massive Multitask Language Understanding
MMT	Multimodal machine translation
MSA	Multimodal sentiment analysis
MoE	Mixture of Experts
MultiRC	MultiSentence Reading Comprehension
NLLB	No Language Left Behind
NLG	Natural language generation
NLP	Natural language processing
NLTK	Natural Language Toolkit
OPT	Open Pre-trained Transformer
OSCAR	Open Super-large Crawled Aggregated coRpus
PaLM	Pathways Language Model
PII	Personally identifiable information
QA	Question answering
RACE	ReAding Comprehension Dataset From Examinations
ReCoRD	Reading Comprehension with Commonsense Reasoning Dataset
ReLU	Rectified Linear Unit
RLHF	Reinforcement Learning from Human
ROOTS	Responsible Open-science Open-collaboration Text Sources
RoPE	Rotary positional embeddings
RTE	Recognizing Textual Entailment
SwiGLU	Swish-Gated Linear Unit
TPU	Tensor Processing Unit
TPV	Total point value
UD	Universal Dependencies
USPTO	United States Patent and Trademark Office
VLI	Video language inference
VQA	Visual question answering
WiC	Word-in-Context
WSC	Winograd Schema Challenge

Abstract

Until recently, natural language processing required a variety of specialized language resources to create functional monolingual and multilingual applications, such as monolingual, bilingual and multilingual corpora, lexical and conceptual resources. Annotated corpora were often needed to enable machine learning techniques in almost all applications within the field. Unprecedented advances in artificial intelligence and the development of large language models, which enable the successful completion of the same and a much wider range of natural language processing tasks, set new standards for the criteria that language resources must meet now and in the near future.

The main objective of the project ***Artificial Intelligence Data Kit 2030 (AID2030)*** is to specify the data kit required to develop computer applications widely known as artificial intelligence in the branch of natural language processing. The breakthrough capacities of language models have been empirically witnessed, but it is difficult to accurately predict the magnitude of future breakthroughs (Way et al., 2022, p. 34). In the conditions of: a) rapid technological development; b) varying degrees of technological support for different European languages, it is not viable to suggest a single static universal kit of text, audio, image, and video data and technologies as it was proposed by the *Basic LAnguage Resource Kit* (Krauwer, 2003).

Based on the existing studies and their in-depth analysis, we propose *Artificial Intelligence data kit 2030* for language understanding, generation, and transformation, set up on a set of criteria to which the data should be adapted depending on the general technological advancement and the specific technology support for different European languages.

1. Introduction

Artificial intelligence (AI) is often regarded as the capacity of computer systems to accomplish tasks inherent for humans and the natural language processing (NLP) is the branch that deals with the human language. According to the highlights of the *Artificial Intelligence Index Report 2023* (Maslej et al., 2023, p. 23), the specific AI topics that continue to dominate research include pattern recognition, machine learning, and computer vision. All three topics are related to natural language processing.

Artificial neural networks, commonly referred to as deep learning, are involved in artificial intelligence as a subset of machine learning. Deep learning algorithms are known to demand huge amounts of training data and computing capacity. These assets have grown more accessible in the recent decade, largely due to the development of complex big data and cloud computing. Significant progress has been made in a range of AI and NLP tasks since the deep learning model Transformer was introduced (Vaswani et al., 2017), and pretrained models based on it have achieved state-of-the-art performance (Qiu et al., 2020; Lin et al., 2022).

Large language models (LLMs) are an AI technology that “understands”, summarizes, generates, and predicts new content using deep learning techniques on massive datasets. Large language models are scaled to more than one trillion parameters (Ren et al., 2023), and it is anticipated that models will become orders of magnitude bigger over the next few years necessitating larger training datasets. It may also be expected that continued performance improvement will result from advancements in training techniques and architectural design. Some approaches produce significant results in training models with a relatively small number of parameters (Touvron et al., 2023), demonstrating a clear relationship between the number of parameters and the size of the datasets: lowering parameters demands larger datasets. The breakthrough capacities of language models have been empirically witnessed,

but it is difficult to accurately predict the magnitude of future breakthroughs (Way et al., 2022, p. 34). However, it is possible to conclude that training or experimenting with larger amounts of data will result in better performance of large language models, and the requirement for the quality and diversity of the data should be on the same scale as its quantity.

The main objective of the project ***Artificial Intelligence Data Kit 2030 (AID2030)*** is to specify the data kit required to develop computer applications widely known as artificial intelligence in the branch of natural language processing. In the conditions of: a) rapid technological development; b) varying degrees of technological support for different European languages, it is not viable to suggest a single static universal kit of text, audio, image, and video data and technologies as it was proposed by the *Basic LAnguage Resource Kit* (BLARK) (Krauwer, 2003).

Based on the existing studies and their in-depth analysis, we propose *Artificial Intelligence data kit 2030* for language understanding, generation, and transformation, set up on a **set of criteria** to which the data should be adapted depending on the general technological advancement and the specific technology support for different European languages. Since simultaneous processing of text, speech, images, and videos is a developing trend, the term "data" rather than the term "language resources" (LR) is used.

The following themes are covered in the document: the Basic Language Resource Kit is introduced in Section 2; Section 3 contains a description of the research methodology; Section 4 offers a survey of the most recent and notable LLMs, demonstrating trends and advancements; Sections 5 and 6 represent surveys on the most significant datasets and benchmarks currently available for LLMs training and evaluation; an overall analysis and specification of the *Artificial Intelligence Data Kit 2030* are presented in Section 7; and a conclusion is presented in Section 8 preceding references and appendices.

2. BLARK: brief overview

Over 20 years ago, ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) collaborated to describe the **BLARK** (Basic LAnguage Resource Kit) concept. BLARK is defined as the bare minimum of resources required to do any precompetitive research or education (Krauwer, 2003). Numerous resources are included in BLARK: (mono- and multilingual) corpora of written and spoken language, mono- and bilingual dictionaries, terminology collections, grammar resources, taggers, morphological analyzers, parsers, speech analyzers, and recognizers, among others. A report describing a (minimum) set of language resources to be made available for as many languages as feasible was developed by ELDA (Evaluations and Language resources Distribution Agency).[1]

The BLARK concept served as inspiration for the creation of equivalent standards for a variety of languages, including Arabic (Maegaard, 2004), Persian (Seraji et al., 2012), seven Central and Southeast European languages (Váradi, 2014), among others. For example, to select a representative set of language resources and tools for the Central and Southeast European languages, a list of general indicators was formulated. The indicators determine the general requirements to which the selection of resources should be subjected. For each indicator, various sets of specific criteria have been established. The general indicators are:

a) For upgraded resources: all selected resources are state-of-the-art specimens of their type for a given language; equally valuable representatives are all included in the selection; etc.

b) For extended/linked resources: the extension of resources provides considerable value to the community, at least on a regional level; the emphasis is on providing building blocks

[1] http://www.blark.org

to the existing tools rather than major restructuring; etc.

c) For resources aligned across languages: no more than one tool of a certain type for each language is used; whenever applicable, the largest set of languages is selected; etc.

The general indicators were combined with the total point value (TPV) (Váradi, 2014, p. 15) evaluating the availability, quality, quantity, and standards of the language resources and tools under selection.[2] The intellectual property rights (IPR) and legal issues were also taken into consideration, promoting the use of open data and following the Creative Commons and Open Data Commons principles.

The approach to outline the set of criteria (quantitative, qualitative, standardization, licensing) for language resources and tools at a certain stage in technological development is admittedly plausible. The formulation of BLARK was one of the factors that fostered the development of language resources and technologies for machine translation, speech processing, language extraction and transformation, and facilitated the measurement of the technological readiness of European languages in 2012 (Rehm and Uszkoreit, 2012).

Until recently, the NLP paradigm required the execution of a pipeline of interdependent modules that could also work alone to arrive at a concrete solution. For example, text categorization was typically preceded by tokenization, sentence splitting, part-of-speech tagging, and lemmatization. Each module was related to specific language resources that are required for training and evaluation. There are numerous well-known examples of these pipelines, including Natural Language Toolkit (NLTK) (Bird et al., 2009), Stanford NLP (Manning et al., 2014), UDpipe (Universal Dependencies) (Straka et al., 2016), and NLP-Cube (Boros et al., 2018).

To some extent, BLARK still matters when it comes to the technological readiness of under-resourced languages, for which there were no language resources or tools for their processing until recently. On the other hand, the unheard-of technological advancement in artificial intelligence and the creation of large language models, which enable the successful completion of the same and a much wider range of tasks in the field of natural language processing, set new requirements for the criteria that language resources must meet at present and in the near future.

3. Methodology of Research

Specifying a data kit designed for artificial intelligence and meeting criteria such as quantity, diversity, quality, flexibility, linking, and standardization is based on:

- An analysis of the most recent and influential applications, services, and resources.
- An analysis of the current credible research, starting with the reports and findings of European Language Equality (ELE) (Agerri et al., 2021; Aldabe et al., 2021; Bērziņš et al., 2022; Backfried et al., 2022; Gomez-Perez et al., 2022; Kaltenboeck et al., 2022) but also including more recent scientific publications, surveys, and artificial intelligence strategies. (Appendix A summarizes the AI development strategies for particular nations or institutions that were not accessible for the corresponding European Language Equality report in 2021.)
- Interactions with representatives from European companies implementing language technology and artificial intelligence. (The findings of a survey seeking information for the use, development, and vision of AI applications are included in the Appendix B.)

Drawing on the listed studies and their in-depth analysis, we propose an AI data kit for language understanding, generation, and transformation, as well as a set of criteria to which the

[2] http://cesar.nytud.hu/deliverables/d2.4-report-on-methodology-and-criteria-for-the/d2.4.report_on_methodology-v1.2.pdf

data kit will be adapted depending on technological advancement and the specific technology support for different languages.

4. Large Language Models

Traditional transfer learning methods use annotated data for supervised training and are focused on a particular task. Since the shift to deep learning, the dominant transfer learning approach has been pre-training methods that use unannotated data for self-supervised training and are applied to various downstream tasks via fine-tuning or few-shot learning (Wang et al., 2022a). Generative Pre-trained Transformer (GPT) (Radford et al., 2018) was the first model that used unidirectional transformers as the backbone for language models, demonstrating the potential for diverse downstream tasks.

The effectiveness of large language models has sparked widespread interest, and a number of attempts have been made to scale them up and explore their performance, for example, HyperCLOVA, a Korean variant of 82B GPT-3 (Kim et al., 2021), CPM-2, a large-scale generative Chinese pre-trained language model (Zhang et al., 2021), Switch Transformers, a mixture of experts model (Fedus et al., 2022), Yuan 1.0, a large-scale Chinese pre-trained language model in zero-shot and few-shot learning (Wu et al., 2021), GLaM, mixture of experts scaling language models (Rae et al., 2022), Gopher, scaling language models (Rae et al., 2022), and so on (Table 1).

Model	Parameters	Architecture	Language
GPT-3 (Brown et al., 2020)	175B	Decoder	English
HyperCLOVA (Kim et al., 2021)	204B	Decoder	Korean
CPM-2-MoE (Zhang et al., 2021)	198B	Encoder–decoder (seq2seq)	Chinese, English
Switch transformers (Fedus et al., 2022)	1751B	Encoder–decoder (seq2seq)	English
Yuan 1.0 (Wu et al., 2021)	245B	Encoder–decoder (unified)	Chinese
GLaM (Du et al., 2022)	1.2T	Encoder	English
Gopher (Rae et al., 2022)	280B	Decoder	English

Table 1: Pre-trained language models (Wang et al., 2022a)

There have already been several surveys on large language models, describing: advances in natural language processing via large pre-trained language models (Min et al., 2021), categorization of existing pre-trained language models based on a taxonomy with four perspectives (Qiu et al., 2020), models for text generation (Li et al., 2021a), transformer-based pre-trained language models (Kalyan et al., 2021), opportunities and risks of foundation models (Bommasani et al., 2021), history of pre-training and breakthroughs (Han et al., 2021), scaling and impact of pre-trained models (Wang et al., 2022a), prompting methods in natural language processing (Liu et al., 2023a), hystory of pre-trained models: from BERT to ChatGPT (Zhou et al., 2023), multimodal pre-trained models (Wang et al., 2023), recent advances of LLMs with introducing the background, key findings, and mainstream techniques (Zhao et al., 2023).

The last survey is accompanied with up-to-date lists for publicly available models, closed-source models, commonly used corpora, library resources, deep learning frameworks, as well as with bibliography for pre-training (data collection, architecture, training algorithms, pre-training on code), adaptation tuning (instruction tuning, alignment tuning), utilization, and capacity evaluation.[3]

Several ground-breaking large language models have emerged recently, pushing the frontiers of what computers can "understand", transform, and generate. An extensive list of such models is presented in the *AI Index 2023 Annual Report* (Maslej et al., 2023). Here are presented some of the most significant language models introduced in 2022 and before May 2023, in an attempt to outline the widest scope of achievements.

4.1. PaLM

PaLM (Pathways Language Model) has been released by Google[4] at April 2022.[5] PaLM is a 540-billion parameter, Transformer-based language model. PaLM was trained on 6144 TPU[6] v4 chips using Pathways, a new machine learning system which enables highly efficient training across multiple TPU Pods (Chowdhery et al., 2022, p. 7). The model demonstrates scaling and achieving state-of-the-art few-shot learning results on hundreds of language understanding and generation benchmarks.

- **Focus**
- Further understanding the impact of scale on few-shot learning.
- Demonstration of the capabilities in language understanding and generation across a number of difficult tasks, such as multi-step mathematical or commonsense reasoning.
- **Training data**

The PaLM pre-training dataset is made up of a high-quality corpus of 780 billion tokens representing a diverse variety of natural language use cases (Chowdhery et al., 2022, pp. 6–7). The dataset is a mixture of multilingual social media conversations (50%); multilingual filtered webpages (27%); books in English (13%), source code obtained from open source repositories on GitHub (5%), Wikipedia articles in many languages (4%), and news articles in English (1%). This dataset is based on the datasets used to train Language Models for Dialog Applications (LaMDA) (Thoppilan et al., 2022) and General Language Model (GLaM) (Du et al., 2022). All three models were trained on exactly one epoch of data (shuffled identically), and the mixing proportions were chosen so as to avoid repeating data in any subcomponent (Chowdhery et al., 2022, p. 6).

- **Languages**

The training mixture includes 124 languages, with English accounting for approximately 78% of the training tokens and only 4 languages represented by more than 10B tokens: German, French, Spanish, and Polish (Chowdhery et al., 2022, p. 73).

- **Approach**

[3] https://github.com/RUCAIBox/LLMSurvey/tree/main

[4] https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

[5] Google released PaLM 2 in May 2023. PaLM 2 is claimed to outperform earlier state-of-the-art LLMs, including PaLM, at complex reasoning tasks such as code and math, categorization and question answering (QA), translation and multilingual proficiency, and natural language production. The combination of compute-optimal scaling, an improved dataset mixture, and model architecture improvements is pointed out as the main reason for this advancement. https://ai.google/discover/palm2

[6] TPU: Tensor Processing Unit

The Pathways system is used to scale the training of a 540-billion parameter language model (Chowdhery et al., 2022, pp. 7–9).

- PaLM 540B is trained over two TPU v4 Pods using data parallelism at the Pod level, while using standard data and model parallelism within each Pod.
- The system, which was the largest TPU configuration documented up until the PaLM release date, allowed for efficient scaling of training to 6144 chips.
- The model FLOPs[7] utilization without rematerialization expenses was 45.7% without self-attention and 46.2% with it. PaLM's analytically estimated hardware FLOPs usage was 57.8%, including rematerialization FLOPs.
- **Evaluation**

Experiments show that with scaling up to the largest model, model performance increased significantly. The PaLM 540B excelled at a variety of highly demanding tasks (Chowdhery et al., 2022, pp. 11–33).

- It was reported for a few-shot performance, achieving state-of-the-art results on 28 out of the 29 most widely evaluated English NLP tasks (including question answering, sentence-completion, reading comprehension, common-sense reasoning, and SuperGLUE tasks) when compared to the best per-task result from any previous large language model.
- PaLM 5-shot outperforms the average performance score of people who were asked to complete the identical tasks on BIG-bench, a benchmark encompassing challenging new language problems.
- Even with a relatively small fraction of non-English data (22%), the 540B model few-shot evaluation results bridge the gap with the prior fine-tuned state of the art in non-English summarization tasks and outperform the prior state of the art in translation tasks.
- **Access**

An unofficial PyTorch implementation of the Transformer architecture based on the PaLM research paper is available on GitHub.[8] It will not scale and is only available for educational use.

4.2. NLLB

NLLB (No Language Left Behind), which was published in August 2022 by Meta,[9] integrates open-source models capable of delivering evaluated, high-quality translations directly between over 200 languages.

- **Focus**
- Datasets and models that narrow the performance gap between low- and high-resource languages.
- State-of-the-art translation models for a wide range of languages.

[7] FLOPs: Floating point operations
[8] https://github.com/lucidrains/PaLM-pytorch
[9] https://ai.facebook.com/research/no-language-left-behind/

- **Training data**

Bitext is available for 148 English-centric and 1,465 non-English-centric language pairs in the dataset. For low-resource languages, parallel data was either selected or constructed. The following components were developed for the construction of over 1.1 billion new sentence pairs of training data for 148 languages: (1) a high-quality language identification system for over 200 languages; (2) a detailed, documented monolingual dataset curation and cleaning pipeline; and (3) a teacher-student-based multilingual sentence encoder training methodology that enables transfer to extremely low-resource languages with minimal supervised bitext (Costa-jussà et al., 2022, pp. 25–46). The total size of the dataset is 450 GB.

- **Languages**

Over 200 languages are supported.

- **Approach**

Several solutions are offered to overcome the lack of publicly available bitext data for many language pairs (Costa-jussà et al., 2022, pp. 25–47).

- The method entails collecting non-aligned monolingual data and then utilizing large-scale data mining (Schwenk et al., 2021) to discover sentences that are likely to be translations of each other in other languages. The solution for translating many languages is to automatically generate translation pairs by pairing sentences from various monolingual resources.
- The 54.5B conditional compute model NLLB-200, based on Sparsely gated Mixture of experts (MoE), was developed, as were some of the smallest models.
- A teacher-student training approach is adopted that allows for the expansion of language coverage to 200 languages and the generation of huge amounts of data, especially for low-resource languages. The overall approach focuses on starting with a massively multilingual sentence encoder teacher model and adapting it to several different low-resource student models (Heffernan et al., 2022).

- **Evaluation**

A benchmarking dataset for machine translation between English and low-resource languages, Flores-200, was developed (Guzmán et al., 2019).

- The performance of over 40,000 different translation directions was evaluated.
- It was reported that the model NLLB-200 outperforms the nearest state-of-the-art by almost +7.3 spBLEU[10] on average – a 44% improvement (Costa-jussà et al., 2022).

- **Access**

All benchmarks, data, scripts, and models are published on GitHub[11] under an MIT (Massachusetts Institute of Technology) license that allows commercial usage, modification, distribution, and private use.

[10] spBLUE is a BLUE (BiLingual Evaluation Understudy) using different tokenization.
[11] https://github.com/facebookresearch/fairseq/tree/nllb

4.3. GLM-130B

GLM-130B (General Language Model) is a bilingual (English and Chinese) pre-trained language model with 130 billion parameters released in August 2022 by the Knowledge Engineering Group at Tsinghua.[12] Instead of employing the GPT-style architecture, the General Language Model algorithm (Du et al., 2022) was used to take advantage of its bidirectional attention and autoregressive blank infilling objective (Zeng et al., 2023).

- **Focus**
- Development of a bilingual, pre-trained dense model with high accuracy on downstream tasks.
- The availability of the model for download and usage on a single server with suitable GPUs (Graphics Processing Units).
- **Training data**

The 95% of the pre-training data includes 1.2 TB English corpora from the Pile dataset (Gao et al., 2020), 1.0 TB Chinese WuDao Corpora,[13] and 250 GB Chinese corpora (including online forums and encyclopedia) crawled from the web. The 5% of the training dataset was intended for Multitask Instruction Pre-Training (MIP) and comprises all prompts for T0 datasets from PromptSource (Bach et al., 2022), and prompts developed for DeepStruct (Deep Structural Prediction) datasets.

- **Languages**

Bilingual support for both English and Chinese.

- **Approach**

The following solutions were used:

- The GLM-130B explores the potential of a bidirectional GLM as its backbone.
- The backbone model is pre-trained over 400 billion tokens on a cluster of 96 NVIDIA DGX-A100 (8×40G) GPU nodes.
- The Rotary positional encoding (Su et al., 2022), DeepNorm layer normalization (Wang et al., 2022b), and Gaussian Error (Hendrycks and Gimpel, 2020) are adapted in training.
- **Evaluation**

For English, it was reported that the GLM-130B model has better performance than the GPT-3 175B Davinci (+5.0%), OPT-175B (Open Pre-trained Transformer) (+6.5%), and BLOOM-176B (+13.0%) on LAMBADA dataset and slightly better performance than the GPT-3 175B (+0.9%) on MMLU (Massive Multitask Language Understanding) benchmark.

For Chinese, it was reported that the GLM-130B model is significantly better than ERNIE TITAN 3.0 260B on 7 zero-shot CLUE (Chinese Language Understanding Evaluation) datasets: +24.26% and 5 zero-shot FewCLUE (Chinese Few-shot Learning Evaluation) datasets: +12.75%.

- **Access**

The GLM-130B model checkpoints, code, training logs, and related toolkits are open through GitHub[14] with Apache License 2.0.

[12] https://keg.cs.tsinghua.edu.cn/glm-130b/posts/glm-130b/
[13] https://www.scidb.cn/en/detail?dataSetId=c6a3fe684227415a9db8e21bac4a15ab
[14] https://github.com/THUDM/GLM-130B

4.4. BLOOM

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) is an open source model with 176 billion parameters developed by hundreds of AI researchers who sought to create a multilingual language model.[15]

- **Focus**
- Providing a large, unbiased, and multilingual model.
- Free access to developed technologies and resources by many different individuals, and companies.
- **Training data**

BLOOM was trained using the ROOTS (Responsible Open-science Open-collaboration Text Sources) (Laurençon et al., 2022) corpus, which contains 498 Hugging Face datasets (Lhoest et al., 2021) comprising 1.6 TB of text covering 46 natural languages and 13 programming languages (BigScience et al., 2023). Altogether, 252 sources were identified, with at least 21 sources per language category. Texts from relevant websites were included to increase the coverage of Spanish, Chinese, French, and English sources.

- **Languages**

46 natural languages and 13 programming languages are covered.

- **Approach**
- BLOOM was trained using a framework for large-scale distributed training – Megatron-DeepSpeed.[16]
- The evaluation of the choice of state-of-the-art LLMs revealed that causal decoder-only models performed better and can be more efficiently converted after pre-training to a non-causal architecture and objective-an approach. Two architectural deviations were adopted in BLOOM (BigScience et al., 2023, pp. 15–16):
  - Instead of adding positional information to the embedding layer, ALiBi Positional Embeddings (Press et al., 2022) directly attenuates the attention scores based on how far away the keys and queries are.
  - BLOOM was trained with an additional layer normalization after the first embedding layer to avoid training instabilities.
- **Evaluation**

BLOOM zero-shot performance on a wide range of natural language processing tasks has been reported as comparable to the performance of similar models (BigScience et al., 2023, pp. 24–42).

- **Access**

The BLOOM Model Card is distributed with the BigScience BLOOM Rail license 1.0 allowing using the model at no charge with some usage restrictions.[17] The source code for BLOOM has been made available under an Apache 2.0 open source license. An online inference API (Application Programming Interface) is also available on the Hugging Face site.

[15] https://bigscience.huggingface.co/blog/bloom
[16] https://github.com/bigscience-workshop/Megatron-DeepSpeed
[17] https://huggingface.co/bigscience/bloom

4.5. LLaMA

LLaMA (Large Language Model Meta AI): a collection of foundational language models ranging from 7B to 65B parameters, is introduced by Meta AI at 24 February 2023.[18] LLaMA 32.5B and 65.2B were trained on 1.4 trillion tokens, while the smallest model, LLaMA 6.7B, was trained on 1 trillion tokens.

- **Focus**
- A series of language models that achieve the best possible performance by training on more tokens than what is typically used (Touvron et al., 2023).
- Providing the research community with the opportunity to access and study smaller and more effective models.
- Top-performing model training using publicly available datasets instead of private or restricted data sources.
- **Training data**

Only publicly available datasets were utilized for training: pre-processed Common Crawl dumps of English webpages from 2017 to 2020; C4 dataset[19] (Colossal Clean Crawled Corpus), a cleaned version of Common Crawl's web crawl for English (Raffel et al., 2020) (15%); public GitHub dataset available on Google BigQuery (4.5%); Wikipedia dumps from the June-August 2022 period, covering 20 languages (4,5%); two English book corpora (4.5%): the Gutenberg Project[20] and the Books3 section of The Pile[21]; ArXiv Latex files[22] representing scientific texts (2.5%); and a dump from Stack Exchange,[23] a website with questions and answers on a variety of topics (2%).

- **Languages**

20 languages with the most speakers, focusing on those with Latin and Cyrillic alphabets, were chosen. However, these languages are presented in the 4,5% Wikipedia part of the training dataset.

- **Approach**

The network is based on the Transformer architecture (Vaswani et al., 2017). Some improvements to the standard Transformer architecture were introduced (Touvron et al., 2023):

- Using the GPT-3 methodology, the stability of training was enhanced by normalizing the input for each transformer sub-layer, instead of normalizing the output.
- Following PaLM methodology, the performance was improved by replacing the ReLU (Rectified Linear Unit) non-linearity function with the SwiGLU (Swish-Gated Linear Unit) activation function.
- Similarly to the GPTNeo approach, absolute positional embeddings were eliminated in favor of Rotary positional embeddings (RoPE) at every network layer.

The model's training speed is increased as a result of:

[18] https://ai.facebook.com/blog/large-language-model-llama-meta-ai/
[19] https://github.com/allenai/allennlp/discussions/5056
[20] https://www.gutenberg.org
[21] https://pile.eleuther.ai/
[22] https://info.arxiv.org/
[23] https://stackexchange.com/

- Efficient causal multi-head attention implementation by not storing the attention weights and not computing masked key/query scores.
- The amount of activations that are recomputed during the backward pass is reduced with checkpointing.
- The computation of activations and the communication between GPUs over the network were overlapped as much as possible.
- **Evaluation**

It was reported that 13b LLaMA outperformed GPT-3, being over 10 times smaller, and 65B LLaMA is competitive with 70B Chinchilla and 540B PaLM. It was shown that state-of-the-art performance can be achieved by training exclusively on publicly available data, without resorting to proprietary datasets (Touvron et al., 2023).

- **Access**

The LLaMA model is released under a noncommercial license which is granted on a case-by-case basis to academic researchers; individuals affiliated with organizations in government, civil society, and academia; and industry research laboratories.

4.6. GPT-4

GPT-4 (Generative Pre-trained Transformer 4) is a large-scale, multimodal model (released by Open AI in March 2023)[24] that accepts image and text inputs and generates text outputs.

- **Focus**
- Creation of a large-scale, multimodal model that accepts image and text inputs and produces text.
- Development of infrastructure and optimization methods that operate predictably across a wide range of scales (OpenAI, 2023).
- **Training data**

The model utilizes publicly available data (such as internet data) and data licensed from third-party providers (OpenAI, 2023).

- **Languages**

As reported by OpenAI, GPT-4 understands and generates text in more languages than its predecessor, GPT-3.5.

- **Approach**

There is no information available on the architecture, model size, hardware, training compute, dataset construction, or training procedures.

- The GPT-4 model is a Transformer-based model that has been pre-trained to anticipate the next token in a document.
- Reinforcement Learning from Human (RLHF) feedback was used to fine-tune the model.

[24] https://openai.com/product/gpt-4

- **Evaluation**

GPT-4 was tested on a diverse set of benchmarks, including quantitative and qualitative evaluation (OpenAI, 2023, pp. 44–55). GPT-4 outperforms humans on most professional and academic exams, most notably scoring in the top 10% on a simulated Uniform Bar Examination, and existing language models on traditional NLP benchmarks without the need for benchmark-specific crafting or additional training protocols.

- **Access**

The model is available in part through ChatGPT Plus[25] and through an API provided via a waitlist.

4.7. ChatGPT

ChatGPT[26] is an artificial intelligence chatbot developed by OpenAI and released in November 2022. It is constructed on top of the foundational large language model GPT-3.5 and its current version ChatGPT Plus – on GPT-4. The pre-training dataset, techniques, and model parameters for GPT-4 are not made available to the general public. Through the use of both supervised and reinforcement learning techniques, the ChatGPT model has been enhanced. A dataset of labeler demonstrations of the required model behavior, containing a set of labeler-written prompts and prompts sent via the OpenAI API, is used to fine-tune GPT-3 using supervised learning. Then, using reinforcement learning based on user input, a dataset of rankings of model outputs is gathered to further improve the supervised model called InstructGPT (Ouyang et al., 2022).

After the release of ChatGPT, nearly 200 research papers focused on it (Liu et al., 2023b) were published at ArXiv. The analysis on these papers shows that ChatGPT has already been applied in a number of application domains, including communication, text classification, text generation, code generation, inferences (logical deductions from known facts or information), and information extraction, transformation, enhancement, and processing. ChatGPT and similar technologies can be employed as writing assistants, personal helpers, and general issue solvers in a wide range of industries, including finance, insurance, e-commerce, mobility, healthcare, and customer service. Numerous ChatGPT use cases have already been used in production environments, and thousands of professional use cases have been quickly developed, implemented, and validated by outside parties.

In March 2023 Cerebras-GPT, a family of open compute-optimal language models scaled from 111M to 13B parameters, was relesaed. For efficient pre-training, the Cerebras-GPT models were trained on the Pile dataset using DeepMind Chinchilla scaling rules (Dey et al., 2023). It was reported that all Cerebras-GPT models have state-of-the-art training efficiency on both pre-training and downstream objectives. Cerebras-GPT models are available on Hugging Face.[27]

4.8. Key Features of Large Language Models as of May 2023

There were 23 important AI language systems produced in 2022, approximately six times as many as the next most popular system category, multimodal systems, according to the Artificial Intelligence Index Report 2023 (Maslej et al., 2023, p. 45).

The same source draws the conclusion that large language models are getting bigger over time; the number of parameters in newly released large language and multimodal models

[25] https://openai.com/blog/chatgpt-plus
[26] https://openai.com/blog/chatgpt
[27] https://huggingface.co/cerebras

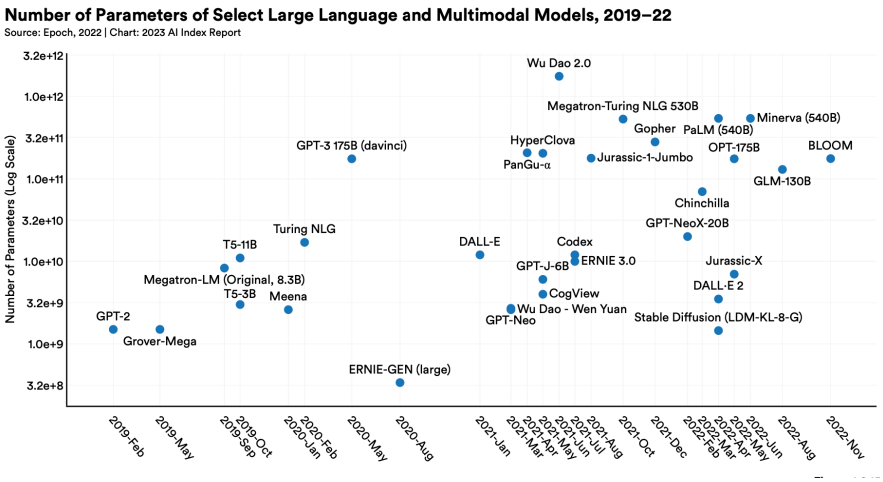

Figure 1: Number of Parameters of Select Language and Multimodal Models (2019–2022) (Maslej et al., 2023, p. 60)

has massively increased. The GPT-2 model, which is the first big language and multimodal model produced in 2019, contained 1.5 billion parameters, compared to the PaLM released in 2022, which has 540 billion parameters, or about 360 times more than GPT-2 (Maslej et al., 2023, p. 60). It is not unexpected that industry advances faster than academics in the field given the resources required for building large language models (compute, data, and expertise). In comparison to the three significant machine learning models generated by academia in 2022, there were 32 significant models produced by industry (Maslej et al., 2023, p. 50). Figure 1 shows the median number of parameters in large language and multimodal models (Maslej et al., 2023, p. 60).

Large language models are a result of powerful technology that excels rapidly at a wide range of language tasks. Without changing the model design, some of the models can address new problems. Widely studied applications of LLMs are document intelligence, which includes sentiment analysis (AlQahtani, 2021), news classification (Ding et al., 2021), anti-spam detection, and information extraction; translation; content creation – creative writing, auto-completion for sentences, paraphrasing, personal decision making, and code generation; virtual assistants, which adopted many applications such as language understanding and generation, and voice recognition (Wang et al., 2022a).

The performance of large language models is improved by increasing the size of the language models (number of parameters). For example, PaLM 540B surpassed the few-shot performance of prior large models, such as GLaM (Du et al., 2022), GPT-3, Megatron-Turing NLG (Natural language generation) (Smith et al., 2022), Gopher (Rae et al., 2022), Chinchilla (Hoffmann et al., 2022), and LaMDA (Hoffmann et al., 2022), that span question answering tasks (open-domain closed-book variant), cloze and sentence-completion tasks, Winograd-style tasks, in-context reading comprehension tasks, common-sense reasoning tasks, SuperGLUE tasks, and natural language inference tasks.[28]

The general performance of large language models is boosted by scaling up the size of the models, the computation used to train them, and the amount of data they're trained on in the appropriate proportions. Scaling laws may properly predict these proportions: larger systems perform increasingly better on a wide range of tasks (Ganguli et al., 2022, p. 1749). It was experimentally determined by training over 400 language models with parameters ranging from 70 million to over 16 billion on 5 to 500 billion tokens that for compute-optimal

[28] https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

training, the model size and the number of training tokens should be scaled equally: **for every doubling of the model size, the number of training tokens should also be doubled** (Hoffmann et al., 2022). Some of the contemporary models are trained on trillions of tokens, and the process of dataset compilation has become a considerable part of the development of large language models. Moreover, the smaller models should have been trained on more tokens to obtain the best performance. For instance, 13B LLaMA outperforms GPT-3 while being more than 10 times smaller, and 65B LLaMA is competitive with 70B Chinchilla and 540B PaLM (Touvron et al., 2023).

The large language models can be fine-tuned with task-specific labeled data on a downstream task (Devlin et al., 2019), however for many tasks, labeled datasets are hardly accessible, and updating the parameters of pre-trained models for various tasks necessitates substantial computing effort.

Some current trends are directed toward generalizing the pre-trained models to many NLP tasks, without fine-tuning them with additional labeled or unlabeled data (Brown et al., 2020). In-context learning entails natural language instruction and/or task demonstrations provided to the language model, after which it can produce the expected output without the need for additional training or updating. Chain-of-Thought is another prompting approach that incorporates intermediate reasoning steps for improving LLM performance on complicated reasoning tasks, including arithmetic reasoning, commonsense reasoning, and symbolic reasoning (Wei et al., 2023).

It was demonstrated that multitask-prompted training can enable strong zero-shot generalization abilities in language models, and such an approach provides an effective alternative to unsupervised language model pre-training (Sanh et al., 2022). The prompting-based techniques have been shown to be successful, but they require **large amounts of unlabeled data for training** and (usually) manual labor to create prompts and select specific tokens to represent class labels.

5. Large Language Datasets

Recent advances in language modelling have shown the effectiveness of training massive models on large text corpora. Each new LLM claims advancements in one or more domains and/or NLP tasks, as well as outperforming some of the other LLMs. Despite the fact that some technologies allow for shorter parts in training datasets for particular domains and/or languages, the growing need for data in language modelling for the majority of languages remains a challenge. Here, we will briefly present some of the widely used and recently compiled language datasets for training large language models.

5.1. CommonCrawl

CommonCrawl[29] creates and maintains an open web crawl data collection. Petabytes of data have been collected by Common Crawl since 2008, including raw web page data, metadata, and text extractions.

Common Crawl is typically used to retrieve subsets of web pages during a given time period. Due to the noisy and low-quality information in web data (Luccioni and Viviano, 2021), it is necessary to perform data cleaning and filtering before usage. There are a number of filtered datasets that are based on Common Crawl: C4 (Raffel et al., 2020), CC-Stories (Trinh and Le, 2018), CC-News (Zhuang et al., 2021), and RealNews (Zellers et al., 2019).

[29] https://commoncrawl.org/

The **C4 dataset**[30] contains over 750 GB of English texts extracted from the Common Crawl web scrape. In addition to considerable deduplication, some heuristics were used to extract only language parts. The dataset was created with the intention of being English-only: any page with a probability of less than 99% being in English was eliminated. **mC4**[31] is a cleaned multilingual version of the C4 dataset (Xue et al., 2021). There are 108 languages available based on the Common Crawl dataset; however, a few instances are romanized variants, written using the Latin script instead of Cyrillic script.

CC-Stories is a dataset for commonsense reasoning and language modeling (not available for download). It was constructed by aggregating documents from the Common Crawl dataset that have the most overlapping n-grams with the questions in commonsense reasoning tasks. The top 1.0% of highest-ranked documents are selected for the new training corpus.

The **CC-News dataset**[32] comprises news articles available at Common Crawl. The version of the dataset was created with an integrated web crawler and news information extractor. It contains 708,241 English-language news stories from January 2017 to December 2019. **RealNews**[33] is a dataset of news articles from Common Crawl. The text and the metadata are extracted from articles in Common Crawl dumps between December 2016 and April 2019. RealNews has a size of 120 GB after deduplication.

OSCAR (Open Super-large Crawled Aggregated coRpus)[34] is a huge multilingual corpus created by language classification and filtering of the Common Crawl corpus. All the data is distributed by 152 languages, both the original and the deduplicated versions of the data are available. The most recent version of OSCAR was released in January 2023, based on the Common Crawl dump from November/December 2022. The size of different languages varies greatly; for example, English texts are 3.4 TB, Chinese texts are 1.4 TB, Russian texts are 1.1 TB, Somali texts are 503 bytes, and Cornish texts are 432 bytes.

5.2. The Pile

The Pile is an 825 GB English text corpus designed for large-scale language model training (Gao et al., 2020). The Pile is constructed from 22 diverse high-quality subsets (already existing or derived from some sources): a Common Crawl–based dataset, Pile-CC, with better quality extractions from the collected html; PubMed – a subset from the PubMed Central – online repository for biomedical articles[35]; Books3 – a dataset of books derived from the contents of the Bibliotik private tracker; ArXiv – preprint research papers predominantly in the domains of Math, Computer Science, and PhysicsGitHub– a large collection of open-source code repositories; Free Law Project[36] – academic studies in the legal realm; Stack Exchange Data Dump[37] – an anonymized set of all user-contributed questions and answers; USPTO Backgrounds – a dataset of background sections from patents granted by the United States Patent and Trademark Office; Wikipedia; PubMed Abstracts – abstracts from 30 million publications in PubMed; PG-19 – a specific Project Gutenberg derived dataset with books from before 1919; the OpenSubtitles dataset (Tiedemann, 2016); the DeepMind Mathematics dataset (Saxton et al., 2019); BookCorpus2 – an expanded version of the BookCorpus (Zhu et al., 2015); the Ubuntu IRC dataset derived from the publicly available chatlogs[38]; EuroParl

30 https://huggingface.co/datasets/c4
31 https://huggingface.co/datasets/mc4
32 https://huggingface.co/datasets/cc_news
33 https://github.com/rowanz/grover/tree/master/realnews
34 https://huggingface.co/datasets/oscar-corpus/OSCAR-2201
35 https://pubmed.ncbi.nlm.nih.gov
36 https://free.law
37 https://archive.org/details/stackexchange
38 https://irclogs.ubuntu.com

(Koehn, 2005) – a multilingual parallel corpus; the YouTube Subtitles dataset – a parallel corpus of human generated closed-captions on YouTube; the PhilPapers dataset – open-access philosophy publications; the NIH Grant abstracts – awarded applications through the ExPorter service;[39] Hacker News[40] – a link aggregator operated by a startup incubator and investment fund; and the Enron Emails dataset (Klimt and Yang, 2004).

97.4% of the Pile dataset is in English. Each part of the Pile is distributed with the specific license depending on the subset. Some syntactic and semantic characteristics of the dataset are documented, such as structural statistics (n-gram counts, language, document sizes), topical distributions, social bias, pejorative content, licensing, etc., as well as the Pile's datasheet (Biderman et al., 2022).

5.3. ROOTS

ROOTS (Responsible Open-science Open-collaboration Text Sources) corpus is a 1.6 TB dataset spanning 59 languages (Laurençon et al., 2022). It was used to train the 176-billion-parameter BigScience large multilingual language model (BLOOM). Both programming languages and natural languages are included in the dataset, which was assembled from two sources: crowdsourced datasets and OSCAR, an online data repository built on Common Crawl.

A combination of monolingual and multilingual language resources were chosen and collectively documented through different initiatives of the BigScience Data Sourcing working group to make up the first part of the ROOTS dataset, which made up 62% of the total dataset size. This produced a collection of 252 sources, with at least 21 sources for each language group taken into consideration. While no topic-based filtering was used, some heuristics based on length and other criteria were used to eliminate some incorrect examples. For the remaining 38% of the final dataset, OSCAR version was chosen based on the Common Crawl snapshot of February 2021 (Ortiz Suárez et al., 2020). Documents that had a very high incidence of words signalling inappropriate content were removed.

The corpus is made up of 46 natural languages from 3 macroareas and 9 language families, including Afro-Asiatic, Austronesian, Basque, Dravidian, Indo-European, Mande, Niger-Congo, and Sino-Tibetan. The majority of the corpus (30.03%) is composed of English, followed by Simplified Chinese (16.16%), French (12.9%), Spanish (10.85%), Portuguese (4.91%), and Arabic (4.6%) (Laurençon et al., 2022, p. 8).

Currently, 223 subsets are accessible upon acceptance of the project's ethical charter.[41] Each subset is released under the license that applies to it. It is possible to access the data cards for each subset.For the remaining subsets, data governance efforts are ongoing to make them as accessible as possible. Tooling code is released under Apache 2.0.[42]

5.4. MassiveText

MassiveText (Rae et al., 2022) is a collection of large English-language text datasets from various sources including web pages, books, news articles, and code. The data pipeline consists of text quality filtering, repetitive text removal, comparable document deduplication, and test-set document removal. MassiveText has 2.35 billion documents, or around 10.5 TB of text. The data from a diverse range of sources was included: web pages (from custom dataset MassiveWeb – 1.9 TB, C4 – 0,75 TB, and Wikipedia – 0.001 TB), books (2.1 TB), news articles (2.7 TB), and code (GitHub – 3.1 TB). The vast majority – 99% – of text in MassiveText is English. Hindi makes up the majority of the non-English text, followed by French, Spanish,

[39] https://reporter.nih.gov/exporter
[40] https://news.ycombinator.com
[41] https://huggingface.co/bigscience-data
[42] https://github.com/bigscience-workshop/data-preparation

German, Italian, Japanese and Chinese. The dataset is not distributed to third parties (Rae et al., 2022, p. 49).

5.5. Reddit

Reddit[43] is a social media for thousands of communities and conversation (more than 13B posts and comments). Highly rated postings are frequently regarded as useful and can be used to generate high-quality datasets, such as OpenWebText.[44] PushShift[45] is another Reddit corpus that has been extracted. It is a real-time updated dataset. Pushshift not only delivers monthly data dumps, but also valuable utility tools to let customers search, summarize, and do investigations on the dataset. Another dataset designed for abstractive summarization is Reddit Webis-TLDR-17.[46] The dataset consists of 3,848,330 posts with an average length of 270 words for content, and 28 words for the summary.

5.6. RedPajama

RedPajama[47] replicates the LLaMA training dataset, which contains over 1.2 trillion tokens. It consists of: five Common Crawl dumps (878 billion tokens), C4 dataset (175 billion), GitHub data (59 billion), Books – a corpus of open books, deduplicated by content similarity (26 billion), ArXiv scientific articles (28 billion), Wikipedia articles (24 billion), StackExchange – a subset of popular websites under StackExchange (20 billion). The RedPajama is licensed under the Apache License, Version 2.0. The datasets are distributed with particular licenses they use. The data pre-processing and quality filters are also available on GitHiub.

5.7. Wikipedia

Wikipedia is an online encyclopedia with a significant number of high-quality articles on a variety of topics. The majority of these articles are written in an expository style (with accompanying references) and span a wide range of languages and fields. In most LLMs, the English-only filtered versions of Wikipedia are commonly used, although Wikipedia is available in a variety of languages, allowing it to be used in multilingual processing.

5.8. Key features of Large Language Datasets as of May 2023

In general, LLMs with a significantly greater number of parameters need a greater volume of training data covering a broader range of content. The quality of the data for pre-training is also of great importance for the model's capacities.

It was already shown that as the parameter scale in the LLM increases, more data is necessary to train the model (Hoffmann et al., 2022). A similar scaling law is also seen in data size with regard to model performance. By conducting extensive experiments, it was further observed that increasing the model size and data size at equal scales can lead to a more compute-efficient model. On the other hand, it was recently demonstrated that with more data and longer training, smaller models may also attain high performance (Touvron et al., 2023). Overall, **the amount of high-quality data required to successfully train models is important**, particularly when increasing or optimizing the model parameters.

[43] https://www.reddit.com
[44] https://skylion007.github.io/OpenWebTextCorpus/
[45] https://pushshift.io
[46] https://huggingface.co/datasets/reddit
[47] https://github.com/togethercomputer/RedPajama-Data

Large part of the datasets used to train LLMs are not publicly available. However, there are several exceptions, such as the Pile (Gao et al., 2020) and RedPadjama which are curated corpora of mostly of English datasets representing various domains; ROOTS (Laurençon et al., 2022) which is a corpus with selected datasets encompassing several domains and languages; C4 (Raffel et al., 2020), mC4 (Xue et al., 2021) and OSCAR (Ortiz Suárez et al., 2020), which are improved versions of Common Crawl and are usually used as part of bigger training datasets.

Most LLMs rely on broad internet data (Chowdhery et al., 2022; BigScience et al., 2023; Touvron et al., 2023), such as websites, webnews, social media, and Wikipedia. On the one hand, obtaining such data is rather simple; nevertheless, the data must be cleaned, screened, and filtered for improper content and redundancies. The Common Crawl corpus is one of the primary sources of pre-training datasets; nevertheless, considerable efforts are invested to extract or re-collect, clean, and filter its data. Wikipedia appears to be utilized with little change; however, it only represents a small portion of the vast training datasets. Furthermore, in order to train LLMs that are adaptable to specific applications, it is necessary to obtain data from relevant sources in order to augment the related information in pre-training data.

The data from different domains or scenarios has distinct linguistic characteristics and contains different semantic knowledge. LLMs can acquire a broad scope of knowledge and a good generalization strength by pre-training on a mixture of datasets from a variety of sources. When combining diverse sources, the distribution of pre-training data should be carefully determined since it is likely to impact LLM performance on downstream tasks (Rae et al., 2022).

The LLMs are usually trained on vast amounts of data from only a few languages, language pairs, and domains as the amount of multilingual data in large language datasets is relatively small. PaLM and BLOOM, which were pre-trained with multilingual datasets, perform well in multilingual tasks such as translation, multilingual summarization and multilingual question answering, and achieve comparable or superior performance to state-of-the-art models fine-tuned on multilingual data.

It is expected that AI advancement will continue to be mostly associated with high-resource conditions, yet the majority of domains and languages, including those in Europe, are under- or low-resourced. The systems should be able to train across several domains and genres as well as cover all European languages (rather than just English or another language with more resources). A wider range of domains and languages should be represented in publicly accessible multilingual datasets so that LLMs with good performance can be developed.

6. Benchmarks

The creation of robust evaluation datasets, or benchmarks, that can gauge advancement in the field, has received a lot of attention in machine learning research. The capacity for evaluation allows for a comparison of various strategies and determines further research and improvement. Three basic types of evaluation tasks for LLMs are defined: language generation, knowledge utilization, and complex reasoning (Zhao et al., 2023, pp. 29–34). First, the three basic types will be briefly discussed, followed by a description of some frequently used evaluation benchmarks the tasks of which can be attributed to any of the listed types.

6.1. Language Generation

Language modeling, as a fundamental aspect of LLMs, aims to predict the next token based on the previous tokens, with a primary focus on basic language understanding and generation (Zhao et al., 2023, p. 29). For example, the LAMBADA dataset (Paperno et al., 2016) is

used to evaluate the modeling capacity of long-range dependencies in text: LLMs have to predict the last word of sentences based on a paragraph of context. **Conditional text generation** (Zhao et al., 2023, p. 29) is concerned with generating texts that satisfy specific task demands, depending on the provided conditions, which commonly include machine translation, text summarization, and question answering. Evaluation benchmarks that are widely used are SuperGLUE (Wang et al., 2019) or MMLU (Hendrycks et al., 2021) but LLMs outperform human raters on some tasks, and this necessitates the development of new benchmarks such as BIG-bench (Srivastava et al., 2023) by collecting currently unsolvable tasks (tasks on which LLMs fail to perform well) or creating more challenging tasks. An example for machine translation evaluation benchmark is Flores-200 (Guzmán et al., 2019), an extension of Flores-101. The Benchmark consists of 3,001 sentences sampled from English-language Wikimedia projects for 204 total languages. Approximately one third of the sentences are collected from each of these sources: Wikinews, Wikijunior, and Wikivoyage. The content is professionally translated into 200+ languages to create Flores-200.

6.2. Knowledge Utilization

Knowledge utilization is the ability to accomplish knowledge-intensive tasks (e.g., commonsense question answering and fact completion) based on supporting factual information (Zhao et al., 2023, p. 31). **Closed-book Question Answering** tasks test the acquired factual knowledge of LLMs from the pre-training dataset, where LLMs should answer the question based on the context without using external information. For example, WebQuestions dataset[48] (Berant et al., 2013) is a question – answering dataset that contains 6,642 question-answer pairs. The original split had 3,778 examples for training and 2,032 testing instances. **Open-book Question Answering** is a task, in which LLMs extract useful information from the external knowledge base or document collections, and then answer the question using that information (Zhao et al., 2023, p. 31). OpenBookQA,[49] for example, is a dataset designed to measure the human understanding of a subject. It comprises 5,957 multiple-choice elementary-level scientific questions (4,957 train, 500 development, 500 test) that assess knowledge of 1,326 key science facts as well as their application to different situations. For training purposes, the dataset provides a mapping from each question to the basic science fact it was designed to evaluate. Answering OpenBookQA questions necessitates broad, common knowledge.

6.3. Complex Reasoning

Complex reasoning is the ability of understanding and utilizing supporting facts or logic in order to reach conclusions or make decisions (Zhao et al., 2023, p. 32). **Knowledge Reasoning** refers to the tasks that rely on logical relations and evidence about factual knowledge to answer the given question. For example, WinoGrande is a large-scale dataset (Sakaguchi et al., 2021) inspired by the original Winograd Schema Challenge. It consists of 44K problems adjusted to improve both the scale and the hardness of the dataset. It was constructed by large-scale crowd sourcing, followed by algorithmic data bias reduction. The dataset is available[50] under Apache license 2.0.

[48] https://worksheets.codalab.org/worksheets/0xba659fe363cb46e7a505c5b6a774dc8a
[49] https://allenai.org/data/open-book-qa
[50] https://github.com/allenai/winogrande

6.4. SuperGLUE

SuperGLUE (General Language Understanding Evaluation)[51] is a benchmark designed to present a more demanding test of general-purpose language understanding for English. SuperGLUE consists of eight language comprehension tasks, a single-number performance metric, and an analysis toolkit that follow the basic design of GLUE (Wang et al., 2019).

GLUE benchmark (Wang et al., 2018) is a framework for evaluating general-purpose language understanding technologies. GLUE is a collection of nine language understanding tasks based on existing public datasets. It also includes private test data, an evaluation server, a single-number target metric, and an additional diagnostic set. With regard to a range of training data volumes, task genres, and task formulations, GLUE was created to offer a general-purpose evaluation of language understanding.

The eight tasks of SuperGLUE are as follows:

BoolQ (Clark et al., 2020) is a question answering task where each example comprises of a short passage and a yes/no question regarding the passage paired with a paragraph from a Wikipedia article containing the answer.

CB (Commitment Bank) (de Marneffe et al., 2019) is a corpus of short texts in which at least one sentence has an embedded clause, each of which is annotated with the author's level of belief in the clause's truth. The Commitment Bank contains 1,200 examples extracted from three corpora of different genres including the Wall Street Journal's news article, the British National Corpus' fiction section, and Switchboard's dialogues.

COPA (Choice of Plausible Alternatives) (Roemmele et al., 2011) is a causal reasoning task. The 1000 handcrafted questions that make up COPA are based on blogs and an encyclopedia of photography-related topics. The task is to choose the alternative that more plausibly has a causal relation with the premise out of the two options for each question.

MultiRC (Multi-Sentence Reading Comprehension) (Khashabi et al., 2018) is a question answering task where each example consists of a context paragraph, a question regarding that paragraph, and a list of possible answers. Each question can have numerous possible valid answers, and each question-answer pair has to be evaluated independently. The examples are from seven domains, including news, fiction, and historical text.

ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset (Zhang et al., 2018) is a multiple-choice question answering task. Each example includes a news article from CNN or the Daily Mail, as well as a Cloze-style question about the article in which one entity is masked out. The system must identify the masked-out entity from a given list of potential entities.

RTE (Recognizing Textual Entailment) dataset (Poliak, 2020) combines several textual entailment datasets and converts them to two-class classifications: entailment and not entailment. Generally speaking, the task is to determine whether the meaning of one sentence can likely be inferred from another.

WiC (Word-in-Context) (Pilehvar and Camacho-Collados, 2019) is a word-sense disambiguation task based on sentence pairs drawn from WordNet, VerbNet, and Wiktionary. The aim is to assess whether a polysemous word used in both sentences is used in the same sense.

WSC (Winograd Schema Challenge) (Levesque et al., 2012) a coreference resolution task with examples that include a sentence with a pronoun and a list of noun phrases from the sentence. Out of the options given, the system must choose the proper pronoun referent.

For some languages similar benchmarks were created, GLUE for Chinese (Xu et al., 2020), Korean (Park et al., 2021), Basque (Urbizu et al., 2022), Bulgarian (Hardalov et al., 2023), French version (Le et al., 2020), an Indonesian version (Koto et al., 2020), a version for Indic languages (Kakwani et al., 2020), Russian SuperGLUE (Shavrina et al., 2020), Slovenian SuperGlue (Žagar and Robnik-Šikonja, 2022). Significant efforts were made to adhere to the

[51] https://gluebenchmark.com

original GLUE's guiding principles, spanning a wide range of domains and tasks and including language-specific characteristics. However, there aren't many other languages for which there are comparable benchmarks.

6.5. BIG-bench

The **BIG-bench** (Beyond the Imitation Game Benchmark) is a benchmark designed to generate difficult challenges for large language models so that they can be tested and their potential predicted (Srivastava et al., 2023). BIG-bench includes more than 200 tasks and human performance metrics gathered by crowdsourcing. The same metrics were used to compare model and human performance against gold labels, which were provided by the original creator of each task. Tasks are diverse, embracing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and beyond. The benchmark tasks are novel, span a wide variety of topics and languages, and are not fully solvable by current models. The tasks are primarily designed to evaluate pre-trained models, without task-specific fine-tuning. The BIG-bench GitHub repository[52] includes: the tasks, benchmark API that supports task evaluation on publicly available models, and evaluation results. Furthermore, BIG-bench Lite is offered as a condensed, canonical subset of tasks, allowing for a quicker evaluation than on the full benchmark.

6.6. MMLU

MMLU(Massive Multitask Language Understanding) benchmark consists of multiple-choice questions in English that encompass various branches of knowledge, such as the social humanities, technology, mathematics, engineering and sciences (Hendrycks et al., 2021). There are 57 tasks in total and the questions (15,908 in total) in the dataset were manually collected from freely available sources online. The benchmark is designed to measure knowledge acquired during pre-training by evaluating models in zero-shot and few-shot settings. It ranges in difficulty from an elementary level to an advanced professional level, and it tests both general knowledge and problem-solving abilities. The benchmark is distributed under the MIT license.[53]

6.7. HELM

HELM (Holistic Evaluation of Language Models) benchmark (Liang et al., 2022) improves the transparency of language models. The vast space of potential scenarios (26 in total) and metrics were taxonomyzed. A broad subset based on coverage and feasibility was selected, and a multi-metric approach was adopted with seven metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency) to analyze specific aspects (e.g., knowledge, reasoning, and disinformation). The dataset can be obtained under the Apache License Version 2.0.[54]

6.8. Key features of Benchmarks as of May 2023

The majority of the current benchmarks, including those for reading comprehension, yes/no reading comprehension, commonsense reading comprehension, and logical reasoning, are designed to test the ability of LLMs to understand the English language. The observations that benchmarking practices are heavily concentrated on a small number of datasets for

[52] https://github.com/google/BIG-bench
[53] https://github.com/hendrycks/test
[54] https://github.com/stanford-crfm/helm

each task and that many benchmark datasets flow between multiple task communities and are leveraged to evaluate progress on tasks for which the data was not explicitly designed (Koch et al., 2021) can be added to this conclusion.

Benchmarks for multitask language understanding test the ability of language models to reason across specialized subject domains. Language benchmarks like GLUE and SuperGLUE are commonly criticized for not testing language models' ability to apply the knowledge they learn across different domains.

An emerging theme in the Artificial Intelligence Index Report 2023 is the observed performance saturation across many popular technical performance benchmarks. This is demonstrated by a comparison between the relative improvement since the inception of the 22 benchmarks (total improvement) and the relative improvement over the previous year. The improvement registered for all but seven of the benchmarks is less than 5%. The median improvement within the last year is 4%, while the median improvement since launch is 42.4% (7 of the benchmarks measure text) (Maslej et al., 2023, p. 114). For instance, superhuman performance was achieved on the common SuperGLUE benchmark less than 18 months after it was produced.

The shortcomings of the current benchmarks are summarized as follows (Srivastava et al., 2023, 5–6): a) Many benchmarks are narrow in scope, focusing mostly on a single or a small number of capabilities where language models have previously shown considerable proficiency; b) Recent language-modeling benchmarks have often had short useful lifespans; c) Many of the benchmarks used today rely on human labeling data collection, which is not done by experts or by the task authors.

7. Artificial Intelligence Data Kit 2030

Unprecedented advances in artificial intelligence and the development of large language models, which enable the successful completion of the same and a much wider range of natural language processing tasks, set new standards for the criteria that language resources must meet now and in the near future.

Until recently, natural language processing required a variety of specialized language resources to create functional monolingual and multilingual applications, such as monolingual, bilingual, and multilingual corpora, lexical, and conceptual resources. Annotated corpora are often needed to enable machine learning techniques in almost all applications within the field, and corpora are augmented with additional annotations to convey more information. These annotations may be about named entities, grammatical structures, or other application-specific annotations (such as those for summarization or question answering). The annotated language resources provide computers with the ability to recognize patterns, train models, and then enhance the underlying algorithms.

A plethora of experiments and papers have repeatedly demonstrated that training large language models on raw large language datasets can aid models in learning general language understanding and generation. Fine-tuning based on pre-trained large language models can improve downstream task impacts while avoiding the need to develop downstream task models from the ground up. The expansion of model scale is aided by the advancement of computer power, the ongoing innovation of training methods, and the availability of large language datasets.

A variety of large language models have been introduced in the previous few years such as dense transformer models (Brown et al., 2020; Rae et al., 2022; Smith et al., 2022; Thoppilan et al., 2022), the most massive of which have exceeded 500 billion parameters (Smith et al., 2022; Chowdhery et al., 2022) or even 1 trillion parameters (Ren et al., 2023). Some of the large language models were trained for about 300 billion tokens (Kaplan et al., 2020; Brown

et al., 2020) or 780 billion tokens (Chowdhery et al., 2022). PaLM 2 was reported to be a 340 billion parameter model that has been trained on 3.6 tokens.[55]

There is a trend toward training larger and larger models since expanding language models has improved the state-of-the-art in many language modeling applications (Hoffmann et al., 2022, p. 3). Large language models, on the other hand, confront a number of challenges, such as their conscious computational needs (Rae et al., 2022; Thoppilan et al., 2022) and the demand for gathering more high-quality training data. **Larger, high quality datasets** are expected to play an important role in any further scaling of language models (Hoffmann et al., 2022, p. 3).

Furthermore, the predicted amount of training data is significantly greater than what is currently used to train large models, which emphasizes the significance of dataset collecting in addition to engineering improvements that allow for model scale. There is a strong suggestion that smaller models should have been trained on more tokens to produce the most performant model. The 70B model Chinchilla was trained with 1.4 tokens, proving the prediction that a 4 times smaller model should be trained with 4 times as many tokens (Hoffmann et al., 2022, p. 2). The 32.5B and 65.2B LLaMA models were trained on 1.4 tokens, while 6.7B LLaMA model was trained on 1 trillion tokens (Touvron et al., 2023). Another example in this direction is 40B Falcon (released in March 2023) trained on 1 trillion tokens.[56]

Although there has been tremendous recent progress, making it possible to train larger and larger models, there are studies which imply that dataset scaling needs to receive more attention: scaling in FLOPs (Hoffmann et al., 2022), in training dataset size (Hoffmann et al., 2022; Geiping and Goldstein, 2022), and in number of parameters (Chung et al., 2022; Fernandes et al., 2023). Scaling to larger datasets would be advantageous when the data is of outstanding quality. This necessitates the careful collection of larger datasets with a strong emphasis on dataset quality.

There is a scarcity of sufficient datasets for all languages, especially for those with limited resources, both in terms of quantity and quality. There is training data available for a few languages that are interesting commercially. However, this is not the case for many (the majority) of the languages, and the available datasets are insignificant compared to English. Many studies focus on cross-lingual transfer learning in an attempt to mitigate the impact of this fact (Schuster et al., 2019; Li et al., 2021b). Although such efforts are beneficial, they seldom result in models with similar performance (as in languages with abundant training data). Additional perspectives might be offered by techniques like MMLM (Multilingual Masked Language Modeling), which has been successfully used to learn multilingual (or cross-lingual) representations of language (Goyal et al., 2021). No Language Left Behind also provides several solutions to overcome the lack of publicly available parallel data for many language pairs (Costa-jussà et al., 2022). The approach involves collecting non-aligned monolingual data and then using large-scale data mining to find phrases that are likely to be translations of each other in other languages (Schwenk et al., 2021). A teacher-student training model is used, which enables for greater language coverage and the creation of huge amounts of data, particularly for low-resource language (Heffernan et al., 2022). To complement technological and algorithmic advancements, it may be beneficial to develop novel approaches that also include users more actively in the generation of datasets for training and evaluation purposes.

In the context of technological advancement and the application of large language models in all fields of NLP, a proposal to adapt the well-known BLARK to the current conditions will be inappropriate. Instead, we propose an ***Artificial Intelligence Data Kit*** aimed at specifying the resources required for pre-training and fine-tuning large language models.

We propose a data-driven approach to ***Artificial Intelligence Data Kit*** design that inte-

[55] https://www.deepmind.com/blog/an-empirical-analysis-of-compute-optimal-large-language-model-training
[56] https://falconllm.tii.ae

grates the best practices of dataset collection with the potential of the latest technologies, allowing fast collection, cleaning, filtering, and automatic metadata extraction. The availability of texts (on the web), rather than a predefined model, determines the dynamic expansion of the datasets, including the growth rate, range of samples, and quantity. The structure of the datasets is maintained by rich metadata organized into a comprehensive classification of categories. The rich metadata description makes it simple to create general as well as domain- and purpose-specific sub-corpora with a fixed structure or specified attributes. Thus, the gist of the approach is that AID2030 design should be centered on **massive amounts of mono- and multilingual (and multimodal) datasets and on providing a detailed metadata description**.

The concept of AI Data Kit is based on several criteria:

7.1. Quantity

Quantity: Massive amounts of dynamically updated data that covers a variety of languages, media types, styles, domains, genres, and topics.

Existing LLMs for which public information is available rely heavily on a mixture of diverse textual resources as the pre-training dataset: web pages, conversation data, books and news, scientific data, and code, which may be categorized as general and specialized data (Zhao et al., 2023, pp. 11–12). It was observed that most LLMs use general data, such as webpages, books, and conversational data, since it is huge, diverse, and relatively easily available, and it helps improve language modeling and generalization abilities. Because of the advent of the internet, many different types of data may be obtained: general and specialized, monolingual and multilingual. Conversation data is part of the pre-training data of many LLMs (Du et al., 2022; Thoppilan et al., 2022; Chowdhery et al., 2022). It can be obtained from publicly available conversations (Baumgartner et al., 2020) or collected from social media.

The specialized data is classified as multilingual data, scientific data, and code; here, data from different domains and genres can also be added. Many LLMs have used multilingual data within their pre-training corpora: NLLB – over 200 languages (Costa-jussà et al., 2022), PaLM – 124 languages (Chowdhery et al., 2022), BLOOM – 46 languages (BigScience et al., 2023), LLaMA – 20 languages (Touvron et al., 2023). Such models provide equivalent or higher performance than state-of-the-art models fine-tuned on the corpus in the target language(s) in multilingual tasks, such as translation, multilingual summarization, and multilingual question answering (Zhao et al., 2023, p. 12).

Domain-specific datasets, whether monolingual, bilingual, or multilingual, enable the creation of domain-specific solutions since businesses require a very specialized vocabulary that is also quickly expanding over time as novel concepts are founded. The necessity of a significant amount of data for certain domains and use cases, as well as the necessity of application-related data, are issues that are closely related to the concept of the quantity of large training datasets.

The data needed for training, as well as solutions for overcoming data scarcity for underresourced languages and domains, are crucial factors in the direction of language equality. Although several machine learning algorithms attempt to reduce the overall reliance on data, datasets will keep their importance for training and, consequently, for the fine-tuning and the evaluation of the quality of the models' performance. Thus, the focus should be on **large amounts of data that cover a variety of languages, media types, styles, domains, genres, and topics**.

The datasets should be **as large as possible** in order to support subsampling, with proportions defined for each subset that might be dynamically changed to maximize the performance of the pre-trained model. Fine-tuning datasets could also be extracted from the large language datasets, the inclusion of subsampling in the pre-training or fine-tuning stage is

part of the overall design of the model training.

7.2. Diversity

Diversity: A wide range of sources: web pages, books, news, patents, code, images, video, and speech records, enabling the models to learn how to produce appropriate performance for different scenarios and applications.

Some models offer competitive results on modalities like language, vision, and multimodal data. Multimodal models can be evaluated and utilized on different downstream tasks, such as image or video captioning – describing the content of an input image or video using a couple of sentences; visual dialogue – talking with humans by holding a meaningful dialog about the visual content; multimodal machine translation (MMT) – translating the source language into a different language based on the paired image (Yang et al., 2019); visual question answering (VQA) – producing an answer provided with an image and a question; video language inference (VLI) – aiming at understanding the video and text multimodal data (Liu et al., 2020); multimodal sentiment analysis (MSA) – attempting to aggregate various homogeneous and/or heterogeneous modalities for more accurate reasoning; and so on (Wang et al., 2023, pp. 22–23). Other applications are: text to image generation (Gafni et al., 2022; Ramesh et al., 2022; Saharia et al., 2022); speech recognition (Shao et al., 2023); text to video generation (Singer et al., 2022).

Datasets with multiple modalities convey more information than unimodal datasets, so machine learning models should improve their predictive performance by processing multiple input modalities. It has been demonstrated how multimodal datasets can benefit the development of different NLP tasks (Garg et al., 2022). The strategy of building large models from all of the data available from many modalities (and languages in current multilingual systems) could be supplemented by the development of smaller models. These smaller models should be trained utilizing the most diverse set of data available, assisting under-resourced languages and domains by leveraging expertise from higher-resourced ones.

As previously noted, pre-training data from various domains or contexts contains unique linguistic properties or semantic knowledge. LLMs can acquire a broad scope of knowledge and a strong generalization capacity by pre-training on a mixture of text data from diverse sources (Wang et al., 2023). When mixing diverse sources, the distribution of pre-training data must be carefully controlled, as it is likely to affect the performance of LLMs on downstream tasks (Rae et al., 2022).

Diversity is a horizontal criterion that applies to all modalities. Large language models have been shown to acquire knowledge in a novel area with relatively small amounts of training data from that domain (Brown et al., 2020). This suggests that by combining a large number of smaller, high-quality, diverse datasets, the model's general cross-domain knowledge and downstream generalization capabilities can be increased (Gao et al., 2020).

Diversity as a criterion for a large language dataset has to be understood as **diversity in the origin** of dataset samples, diversity in their **content** (thematic domains and genres), **language diversity**, and **diversity in coverage cultural aspects**. Such an understanding of diversity is in compliance with the European Language Equality concept for equal technological support for all European languages.

7.3. Quality

Quality: Clean and filtered data with no re-duplication or subsumption of samples and no toxic or biased content.

Large datasets that have been scraped from the web may be corrupted during the extraction, may be the result of machine translation and AI applications, and may include offensive

language, biases, or personal data. The quantity (if not the frequency) of such information grows as datasets get bigger, which makes dataset evaluation important.

Regardless of data modality, several key indicators can be utilized to assess data quality: **technical characteristics** (no data interrupted, broken, or lacking content), **uniqueness** (no data repetition), **non-toxic content** (no toxic data, biases, or personally identifiable information – PII), and **human-generated content**.

Existing work generally employs two ways to remove low-quality data from the collected data: classifier-based and heuristic-based (Wang et al., 2023). The first method involves training a selection classifier on high-quality data and then using it to identify and filter out low-quality data (Brown et al., 2020; Chowdhery et al., 2022). Several studies (BigScience et al., 2023; Rae et al., 2022) employ heuristic-based approaches to eliminate low-quality texts through a set of well-designed rules, such as language-based filtering, metric-based filtering (evaluation metrics about the machine-generated data), statistic-based filtering (e.g., the punctuation distribution, symbol-to-word ratio, sentence length, etc.), and keyword-based filtering (unuseful elements in the text, such as HTML tags, hyperlinks, and offensive words).

It has been found that duplicated data in a corpus would reduce the diversity of language models, which may cause the training process to become unstable and thus affect the model's performance (Hernandez et al., 2022). Therefore, it is necessary to deduplicate the pre-training corpus (Wang et al., 2023). Deduplication can be conducted at several granularities, including sentence-level, document-level, and dataset-level. Low-quality sentences with repeated words and phrases should be deleted since they may promote repetitive patterns in language modeling. Existing studies (Lee et al., 2022) largely rely on the overlap ratio of surface features (e.g., words and n-grams overlap) across documents to detect and eliminate duplicate documents with similar contents.

The pre-training data may consist user-generated content holding sensitive or personal information, which raises the potential for privacy violations. As a result, personally identifiable information must be removed from the pre-training dataset. A direct and effective technique is to use rule-based methods, such as keyword spotting, to discover and eliminate personally identifiable information, such as names, addresses, and phone numbers (Laurençon et al., 2022). Other techniques such as machine learning are also applied.

Data quality dimensions commonly include, but are not limited to, **completeness, validity, timeliness, consistency, and integrity** (Sebastian-Coleman, 2013). Overall, the quality of a language technology application is (frequently) determined by the quality of the underlying or utilized data.

7.4. Structure

Structure: Data and metadata organized in a conceptual graph, allowing the extraction of different datasets, respectively, for languages with excellent, good, moderate, fragmentary, and week or no support for language technologies.

The structure of the dataset is ensured by rich metadata, organized into a graph representation of categories. The rich metadata description makes it simple to create generic, domain- and purpose-specific sub-corpora with a set structure or predetermined characteristics. As a further advantage, graph representation allows flexible extension with new relations and categories and shows where merging or splitting categories is permissible. For example, it is possible to merge the metadata with a database of books' descriptions, allowing the automatic assignment of publishing dates or obtaining translations in different languages. Graph representation will increase interoperability – the capacity to integrate different dataset components so they may be utilized separately or in combination without the need for further modifications and filters.

The metadata describes the attributes of the samples in the dataset. The importance of

metadata and the necessity for as much detail as possible in order to establish the relevance of a particular resource were already emphasized with regard to corpora (Burnard, 2005). The categorization can be used as a baseline description of the textual data in the proposed framework:

- **Editorial**: Information (regarding data samples) with respect to their original source (title, author, creation date, and so on). Here information about language, direction of translation (if applicable), etc. can be included.
- **Descriptive**: Classificatory information for the content and context of the data, such as type, modality, domain, style, and genre. Such information should most likely be automatically allocated.
- **Administrative**: Information regarding the samples and the dataset, such as availability, revision status, copyright information, rights management, and license agreements.
- **Statistical**: Number of tokens, words, domain-specific words, sentences, size, etc.

A detailed description of metadata categories and their values is offered by the European Language Grid.[57]

To increase transparency and accountability in the machine learning community, the concept of datasheets for datasets was proposed. Each dataset should be accompanied by a datasheet that details its purpose, creation, composition, intended uses, distribution, maintenance, and other details (Gebru et al., 2021). Although a template for a dataset datasheet is also proposed,[58] the datasheets themselves contain material similar to that for a scientific article since they permit open-ended responses to questions about the development, organization, licensing, etc. of datasets.

The **technological advancement** in the field of AI NLP is difficult to predict since: the quick pace of technological development; the lack of open access to some of the innovative training data, algorithms, and models; and the non-availability of comprehensive regulations of data use. All of these factors simply offer the chance to sketch out anticipated trends when creating appropriate training datasets and targeting novel applications for language analysis, generation, and transformation based on natural language understanding and generation.

The process of gathering, processing, and selecting language and multimodal data is not simple. The procedure entails not only the collection of disparate data, but also its proper documentation, filtering, and detoxification. To experiment with datasets of varying composition and scale, their metadata must include a minimal set of categories whose values contain indicative information, such as the source, time of creation, originality, language, size, thematic domain, genre, purpose, and quality of the resource. The four main activities: **collection, filtering, detoxification, and documentation of data, require serious efforts and should remain the focus for all European languages, regardless of their current technological support**.

8. Conclusions

Massive amounts of data are necessary for the efficient and productive operation of AI based businesses and organizations. Institutions, companies, and countries that make investments in data collection, organization, and storage will benefit in the future. The AI data kit will facilitate:

[57] https://s3.dbl.cloud.syseleven.net/dev-cms/s3fs-public/2021-11/ELG-Deliverable-D2.3-final.pdf
[58] https://www.overleaf.com/latex/templates/datasheet-for-dataset-template/jgqyyzyprxth.pdf

- **Research advance**: fostering data sharing; focusing on trustworthy, standardized, and interoperable data; enabling further better modeling of multimodal and multilingual environments, and showing how modalities can enrich one another.
- **Business driving**: pouring data into developing AI, enabled by concurrent increases in high-quality data, computing capability, and high-speed communication links. Small and medium-sized businesses that lack the funds to invest in such resources are notably helped by the targeted provision of multilingual, multimodal data. The findings of a non-representative survey seeking information from European companies on the use, development, and vision of AI applications are included in Appendix B.
- **Advancement of society**: AI applications, in combination with other technologies, will benefit almost every area of daily life, including tailored learning, care for people with special needs, enhanced communication, and better conditions for work, rest, and entertainment.
- **Policy making**: Although many future AI advancements will be available on a global scale, there could be disproportionate differences for nations that support, develop, and adopt AI. If current legal frameworks were changed to require that all unprotected language-related data become available, a considerable advancement may be realized. The AI data kit will clearly demonstrate the demand for investments by displaying the **necessity of diverse data collections for a variety of applications and particular languages**. The interested parties will have the necessary information for the current situation as well as the next steps required to achieve specific AI solutions.

The unprecedented technological advancements in artificial intelligence and the development of large language models, which enable the successful completion of the same and a much broader range of tasks as “traditional” natural language processing, have established new requirements for the criteria that language resources must meet. To some extent, BLARK is still relevant when it comes to the technological readiness of under-resourced languages, for which language resources or tools for their processing are not available or are available in limited scope.

As large language models become a standard in AI NLP, the language resources that utilize the whole process of the development of LLMs (training large language datasets, fine-tuning resources, and evaluation benchmarks) should meet new standards both for well-resourced and under-resourced languages. The report ***Artificial Intelligence Data Kit 2030*** is regarded as one of the steps in this direction.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. Deliverable D1.2 Report on the State of the Art in Language Technology and Language-centric AI, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Itziar Aldabe, Georg Rehm, German Rigau, , and Andy Way. Deliverable D3.1 Report on existing strategic documents and projects in LT/AI, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Arwa S. M. AlQahtani. Product Sentiment Analysis for Amazon Reviews. *International Journal of Computer Science and Information Technology (IJCSIT)*, 13(3):15–30, June 2021. URL https://aircconline.com/ijcsit/V13N3/13321ijcsit02.pdf.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-demo.9.pdf.

Gerhard Backfried, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz. Deliverable D2.14 Technology Deep Dive – Speech Technologies, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_14__Speech__Technologies.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. *ArXiv*, 2020. URL https://arxiv.org/pdf/2001.08435v1.pdf.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1160.pdf.

Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the Pile, 2022. URL https://arxiv.org/pdf/2201.07311.pdf.

BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari,

Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. BLOOM: A 176b-Parameter Open-Access Multilingual Language Model. *ArXiv*, 2023. URL https://arxiv.org/pdf/2211.05100.pdf.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009. URL https://www.nltk.org/book/.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak

Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Tiberiu Boros, Stefan Daniel Dumitrescu, and Ruxandra Burtica. NLP-Cube: End-to-End raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL https://aclanthology.org/K18-2017.pdf.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Lou Burnard. Metadata for corpus work. *Developing linguistic corpora: a guide to good practice*, 2005. URL http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf.

Aivars Bērziņš, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, and Andy Way. Deliverable D2.13 Technology Deep Dive – Machine Translation, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_13__Machine_Translation_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways. *ArXiv*, 2022. URL https://arxiv.org/pdf/2204.02311.pdf.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. *ArXiv*, 2022. URL https://arxiv.org/pdf/2210.11416.pdf.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 07 2020. ISSN 2307-387X. URL https://aclanthology.org/2020.tacl-1.30.pdf.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv*, 2022. URL https://arxiv.org/pdf/2207.04672.pdf.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124, Jul. 2019. URL https://semanticsarchive.net/Archive/Tg3ZGI2M/Marneffe.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.pdf.

Nolan Dey, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster. *ArXiv*, 2023. URL https://arxiv.org/pdf/2304.03208.pdf.

SiYu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-Doc: A Retrospective Long-Document Modeling Transformer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2914–2927, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.227.pdf.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/du22c.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23:1–39, 2022. ISSN 1533-7928. URL https://jmlr.org/papers/volume23/21-0998/21-0998.pdf.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling Laws for Multilingual Neural Machine Translation, 2023. URL https://arxiv.org/pdf/2302.09650.pdf.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors, 2022. URL https://arxiv.org/pdf/2203.13131.pdf.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. Predictability and Surprise in

Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022. URL https://dl.acm.org/doi/abs/10.1145/3531146.3533229.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*, 2020. URL https://arxiv.org/pdf/2101.00027.pdf.

Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.738.pdf.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *Commun. ACM*, 64(12):86–92, nov 2021. ISSN 0001-0782. URL https://arxiv.org/pdf/1803.09010.pdf.

Jonas Geiping and Tom Goldstein. Cramming: Training a Language Model on a Single GPU in One Day, 2022. URL https://arxiv.org/pdf/2212.14034.pdf.

Jose Manuel Gomez-Perez, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiljevs, and Teresa Lynn. Deliverable D2.15 Technology Deep Dive – Text Analytics, Text and Data Mining, NLU, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_15__Text_Analytics_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-Scale Transformers for Multilingual Masked Language Modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online, August 2021. Association for Computational Linguistics. URL https://arxiv.org/pdf/2105.00572.pdf.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1632.pdf.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021. ISSN 2666-6510. URL https://www.sciencedirect.com/science/article/pii/S2666651021000231.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Ves Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark. *ArXiv*, 2023. URL https://arxiv.org/pdf/2306.02349.pdf.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.154.pdf.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), 2020. URL https://arxiv.org/pdf/1606.08415.pdf.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/pdf/2009.03300.pdf.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling Laws and Interpretability of Learning from Repeated Data. *ArXiv*, 2022. URL https://arxiv.org/pdf/2205.10487.pdf.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. *arXiv*, 2022. URL https://arxiv.org/pdf/2203.15556.pdf.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.445.pdf.

Martin Kaltenboeck, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg. Deliverable D2.16 Technology Deep Dive – Data, Language Resources, Knowledge Graphs, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_16__Data_and_Knowledge_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and S. Sangeetha. AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing. *ArXiv*, abs/2108.05542, 2021. URL https://arxiv.org/pdf/2108.05542.pdf.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *ArXiv*, abs/2001.08361, 2020. URL https://arxiv.org/pdf/2001.08361.pdf.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1023.pdf.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.274.pdf.

Bryan Klimt and Yiming Yang. The Enron Corpus: A New Dataset for Email Classification Research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30115-8.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *ArXiv*, 2021. URL https://arxiv.org/pdf/2112.01716.pdf.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.pdf.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.66.pdf.

Steven Krauwer. The Basic Language Resource Kit (BLARK) as the first Milestone for the Language Resources Roadmap. In *Proceedings of 2nd International Conference on Speech and Computer (SPECOM2003)*, 2003. URL http://www.elsnet.org/dox/krauwer-specom2003.pdf.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://arxiv.org/pdf/2303.03915.pdf.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.302.pdf.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating Training Data Makes Language Models Better. *ArXiv*, 2022. URL https://arxiv.org/pdf/2107.06499.pdf.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy, 2012. ISBN 978-1-57735-560-1. URL https://cs.nyu.edu/faculty/davise/papers/WSKR2012.pdf.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.pdf.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. Pretrained Language Model for Text Generation: A Survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4492–4499. International Joint Conferences on Artificial Intelligence Organization, 8 2021a. URL https://www.ijcai.org/proceedings/2021/0612.pdf. Survey Track.

Zuchao Li, Kevin Parnow, Hai Zhao, Zhuosheng Zhang, Rui Wang, Masao Utiyama, and Eiichiro Sumita. Cross-lingual Transferring of Pre-trained Contextualized Language Models. *ArXiv*, 2021b. URL https://arxiv.org/pdf/2107.12627.pdf.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models. *ArXiv*, 2022. URL https://arxiv.org/pdf/2211.09110.pdf.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 3:111–132, 2022. ISSN 2666-6510. URL https://www.sciencedirect.com/science/article/pii/S2666651022000146.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. VIOLIN: A Large-Scale Dataset for Video-and-Language Inference. *ArXiv*, 2020. URL https://arxiv.org/pdf/2003.11618.pdf.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9), Jan 2023a. ISSN 0360-0300. URL https://dl.acm.org/doi/pdf/10.1145/3560815.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *ArXiv*, 2023b. URL https://arxiv.org/pdf/2304.01852.pdf.

Alexandra Luccioni and Joseph Viviano. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-short.24.pdf.

Bente Maegaard. The NEMLAR project on Arabic language resources. In *Proceedings of the 9th EAMT Workshop: Broadening horizons of machine translation and its applications*, Malta, April 26–27 2004. European Association for Machine Translation. URL https://aclanthology.org/2004.eamt-1.15.pdf.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL https://aclanthology.org/P14-5010.pdf.

Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. The AI Index 2023 Annual Report, April 2023. URL https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ArXiv*, 2021. URL https://arxiv.org/pdf/2111.01243.pdf.

OpenAI. GPT-4 Technical Report. *ArXiv*, 2023. URL https://arxiv.org/pdf/2303.08774.pdf.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.156.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. *ArXiv*, 2016. URL https://arxiv.org/pdf/1606.06031.pdf.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. KLUE: Korean Language Understanding Evaluation. *ArXiv*, 2021. URL https://arxiv.org/pdf/2105.09680.pdf.

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1128.pdf.

Adam Poliak. A Survey on Recognizing Textual Entailment as an NLP Evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.eval4nlp-1.10.pdf.

Ofir Press, Noah A. Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. *ArXiv*, 2022. URL https://arxiv.org/pdf/2108.12409.pdf.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, Sep 2020. doi: 10.1007/s11431-020-1647-3.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling Language Models: Methods, Analysis and Insights from Training Gopher. *ArXiv*, 2022. URL https://arxiv.org/pdf/2112.11446.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/volume21/20-074/20-074.pdf.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, 2022. URL https://arxiv.org/pdf/2204.06125.pdf.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, Alexander Podolskiy, Grigory Arshinov, Andrey Bout, Irina Piontkovskaya, Jiansheng Wei, Xin Jiang, Teng Su, Qun Liu, and Jun Yao. PanGu-Sigma: Towards Trillion Parameter Language Model with Sparse Heterogeneous Computing. *ArXiv*, 2023. doi: 10.48550/arXiv.2303.10845. URL https://arxiv.org/pdf/2303.10845.pdf.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*. Stanford University, 2011.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv*, 2022. URL https://arxiv.org/pdf/2205.11487.pdf.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Commun. ACM*, 64(9):99–106, Aug 2021. ISSN 0001-0782. URL https://dl.acm.org/doi/pdf/10.1145/3474381.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/pdf/2110.08207.pdf.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing Mathematical Reasoning Abilities of Neural Models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/pdf?id=H1gR5iR5FX.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1380.pdf.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.115.pdf.

Laura Sebastian-Coleman. *Measuring Data Quality for Ongoing Improvement*. Elsevier, 2013. ISBN 978-0-12-397033-6. doi: https://doi.org/10.1016/C2011-0-07321-0.

Mojgan Seraji, Beáta Megyesi, and Joakim Nivre. A Basic Language Resource Kit for Persian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2245–2252, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/338_Paper.pdf.

Hang Shao, Wei Wang, Bei Liu, Xun Gong, Haoyu Wang, and Yanmin Qian. Whisper-KDQ: A Lightweight Whisper via Guided Knowledge Distillation and Quantization for Efficient ASR. *ArXiv*, 2023. URL https://arxiv.org/pdf/2305.10788.pdf.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.381.pdf.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data. *ArXiv*, 2022. URL https://arxiv.org/pdf/2209.14792.pdf.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, 2022. URL https://arxiv.org/pdf/2201.11990.pdf.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew D. La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakacs, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C'esar Ferri Ram'irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Gonz'alez, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou, Jan Koco'n, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng' Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar

Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Saleem Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew Leavitt, Matthias Hagen, M'aty'as Schubert, Medina Baitemirova, Melissa Arnaud, Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pi-Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ram'on Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J. Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, S. Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler O'Brien Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, W Vossen, Xiang Ren, Xiaoyu Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yuntao Bai, Zachary Seid, Zhao Xinran, Zhuoye Zhao, Zi Fu Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu, Sahib Singh, and Uri Shaham. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 5, 2023. URL https://openreview.net/pdf?id=uyTL5Bvosj.

Milan Straka, Jan Hajič, and Jana Straková. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1680.pdf.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *ArXiv*, 2022. URL https://arxiv.org/pdf/2104.09864.pdf.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen

Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language Models for Dialog Applications. *ArXiv*, 2022. URL https://arxiv.org/pdf/2201.08239.pdf.

Jörg Tiedemann. Finding Alternative Translations in a Large Corpus of Movie Subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1559.pdf.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, 2023. doi: 10.48550/arXiv.2302.13971. URL https://arxiv.org/pdf/2302.13971.pdf.

Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847, 2018.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. BasquegLUE A Natural Language Understanding Benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.172.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Tamás Váradi. Serving Multilingual Europe: The CESAR Project. In Svetla Koeva, editor, *Language Resources and Technologies for Bulgarian*, page 9–28, Sofia, 2014. Professor Marin Drinov Publishing House of the Bulgarian Academy of Sciences. ISBN 978-954-322-797-6.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-5446.pdf.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-Trained Language Models and Their Applications. *Engineering*, 2022a. ISSN 2095-8099. URL https://www.sciencedirect.com/science/article/pii/S2095809922006324.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers, 2022b.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey. *ArXiv*, 2023. URL https://arxiv.org/pdf/2302.10035.pdf.

Andy Way, Georg Rehm, Jane Dunne, Maria Giagkou, José Manuel Gomez-Perez, Jan Hajič, Stefanie Hegele, Martin Kaltenböck, Teresa Lynn, Katrin Marheinecke, Natalia Resende, Inguna Skadiņa, Marcin Skowron, Tea Vojtěchová, and Annika Grützner-Zahn. D2.18 Report on the state of Language Technology in 2030, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/05/ELE___Deliverable_D2_18__Report_on_State_of_LT_in_2030_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangang Luo, Liang Xu, and Xuanwei Zhang. Yuan 1.0: Large-Scale Pre-trained Language Model in Zero-Shot and Few-Shot Learning. *ArXiv*, abs/2110.04725, 2021. URL https://arxiv.org/pdf/2110.04725.pdf.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.419.pdf.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.41.pdf.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. Visual Agreement Regularized Training for Multi-Modal Machine Translation. *ArXiv*, abs/1912.12014, 2019. URL https://arxiv.org/pdf/1912.12014.pdf.

Aleš Žagar and Marko Robnik-Šikonja. Slovene SuperGLUE Benchmark: Translation and Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2020.coling-main.419.pdf.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against Neural Fake News. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. URL https://proceedings.neurips.cc/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/pdf?id=-Aw0rrrPUF.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *ArXiv*, 2018. URL https://arxiv.org/pdf/1810.12885.pdf.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. CPM-2: Large-scale cost-effective pre-trained language models. *AI Open*, 2:216–224, 2021. ISSN 2666-6510. URL https://www.sciencedirect.com/science/article/pii/S2666651021000310.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models. *ArXiv*, 2023. URL https://arxiv.org/pdf/2303.18223.pdf.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *ArXiv*, 2023. URL https://arxiv.org/pdf/2302.09419.pdf.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *ArXiv*, 2015. URL https://arxiv.org/pdf/1506.06724.pdf.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. RoBERTa: A Robustly Optimized BERT Pretraining Approach with Post-training, August 2021. URL https://aclanthology.org/2021.ccl-1.108.pdf.

A. Appendix: Strategies and Visions about Artificial Intelligence (2022–2023)

Over the past year, several national strategies have been updated, and a number of reviews related to artificial intelligence have been published. The lists and references presented here upgrade the ones published in ELE's Report on Existing Strategic Documents and Projects in LT/AI (Aldabe et al., 2021). The Federal Ministry of the Republic of Austria, the French Science and Society Association TRACES, the Government of Ireland, and Japan presented strategies or visions for the development of artificial intelligence until 2030, which include the development of artificial intelligence applications for the purposes of various professional activities, such as the integration of artificial intelligence in science, engineering, medicine, business, education, and the humanities.

For example, the Artificial Intelligence Mission of Austria until 2030 is to maximize the benefits of AI for citizens. Austria will invest in various future fields, such as research and innovation, infrastructure for industrial leadership, qualification, and training. Artificial intelligence in Ireland is one of six priority programs for which more than €200 million has been earmarked since 2018. Japan's 2030 vision is to become a frontrunner in the use of AI in real-world industries, resulting in increased economic competitiveness.

A number of research groups and institutions provide visions and plans for the development and implementation of artificial intelligence in personal and professional life. The Australian University of Queensland presents analysis and guidance on the use and adoption of AI; the United Nations Educational, Scientific, and Cultural Organization presents a quick guide to ChatGPT and artificial intelligence in higher education; etc.

There are some documents analyzing the policy of using artificial intelligence as well as the moral application of artificial intelligence, rights, and legal compliance (Artificial intelligence governance and human rights, produced by the Royal Institute of International Affairs in London, or Artificial Intelligence and Education: A Critical View Through the Lens of Human Rights, Democracy, and the Rule of Law, published by the Council of Europe, and so on).

As a conclusion about the development guidelines, more **ethical discussions, language models, and AI regulation** are the common themes in (national) AI strategies and visions. There is a trend for collaboration to develop systems and processes that will **significantly advance data trust and professionalize the data space**, with the goal of developing standards for data and organizational data practices that have the potential to create a step change in **data sharing across organizational and geographical boundaries**.

Organization	Title	Year
Oxfords Insights	Government AI Readiness Index 2022	2022
The Observer Research Foundation	G20.AI National Strategies, Global Ambitions	2022
Stanford University	Implementation Challenges to Three Pillars of America's AI Strategy	2022
TRACES Association, France	TRACES – In 2030, Artificial Intelligence Will Visit Museums?	2022
AI Now Institute, USA	AI Now 2023 Landscape: Confronting Tech Power	2023
Department for Science, Innovation and Technology, UK	A pro-innovation approach to AI regulation	2023
National Institute of Standards and Technology, USA	Artificial Intelligence Risk Management Framework (AI RMF 1.0)	2023
Open Data Institute	The ODI five year strategy 2023–2028	2023
Rackspace Technology	The 2023 AI and Machine Learning Research Report	2023
The Royal Institute of International Affairs, London	AI governance and human rights	2023
Stanford University	Artificial Intelligence Index Report	2023
The United Nations Educational, Scientific and Cultural Organization	ChatGPT and artificial intelligence in higher education: quick start guide	2023
The University of Queensland, Australia	Navigating AI – Analysis and guidance on the use and adoption of AI	2023
The University of Queensland, Australia	Trust in Artificial Intelligence	2023
UNECE	White Paper on the use of Artificial Intelligence in Trade Facilitation	2023

Table 2: Institutional AI Strategies and/or Visions (2022–2023)

Country	Title	Year
Africa	Responsible AI in Africa – Challenges and Opportunities	2022
Council of Europe	Artificial Intelligence And Education	2022
EU	Regulation 2022/868 OF the EUP and the Council on European data governance and amending Regulation 2018/1724 (Data Governance Act)	2022
Ireland	Impact 2030 – Ireland's Research and Innovation Strategy	2022
Japan	Japan AI Strategy 2022	2022
Sweden	AI Vision White paper	2022
UK	The UK's AI Strategy: Where Are We Now?	2022
US	Strengthening and Democratizing the U.S. AI Innovation Ecosystem	2023

Table 3: National and Regional AI Strategies and/or Visions (2022–2023)

B. Appendix: Survey *Artificial Intelligence Data Kit*

A non-representative survey (named after the project's name) was planned and conducted among representatives of companies that create or utilize applications in the field of AI NLP.

The purpose of the study is to outline the industrial sectors in which some applications of AI NLP have already entered, the expectations for AI development in individual industrial sectors and in the various stages of creation and sale of goods, as well as the vision for the future of AI NLP. The survey was sent exclusively to the European companies that have already shown interest in AI NLP.

Survey was spread among representatives of 258 companies from 26 countries, including Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Lithuania, Malta, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Ukraine, and the United Kingdom.

The sector distribution of the selected companies is as follows: Computer software – more than 100 companies; data processing services, financial services, healthcare, and technology – between 20 and 40 companies; e-commerce, marketing and advertising, telecommunications, transportation, travel and tourism, supply chain and logistics, insurance, and customer service – below 20 companies.

Among main areas of operation of respondents computer software dominates with 32,3%, followed by data processing services with 29%. Telecommunication and financial services are less represented with 6,5%, the same counts for healthcare and marketing and advertising with 3,8%. The result, to some extent, corresponds with the distributions of domain areas among target companies (Figure 2).

The selected solutions based on language technologies are: customer support – 26,7%, decision support systems – 20%, dynamic data modeling – 11,1%, multilingual analysis – 15,6%, and virtual agents – 17,8% (Figure 3).

The range of technologies already employed reflects diversity in responses. Conversational intelligence is selected by 9% of the respondents, data analytics (monitoring, trends, prediction-making, etc.) by 15,7%, data transformation (extraction, summarization, translation, speech to text, etc.) by 13,5% and deep learning by 15,7%. Under 10% are the selections of emotion artificial intelligence, event detection, image and video processing, intent analysis, natural language understanding, sentiment analysis and speech processing (Figure 4).

Most of the companies (65%) support between one and five languages, 10% – only one, and 25% – more than five languages (Figure 5). 45.5% of the companies adjust already available solutions, 36.4% develop in-house solutions, and 18.2% buy ready-to-use solutions (Figure 6).

Further benefit from AI development is seen in automated helpdesk – 20,4%, customer interactions – 14,3% and marketing – 12,2%. Other AI developments, such as document management, business-to-business interactions, logistics, onboarding new customers, purchase and payment management, robotic process automation, supply chain management and workforce management remain under 10% (Figure 7). It is noteworthy that the interaction with customers occupies an important part of the expectations for the future introduction of AI technologies in business, which is also in line with the results of a significantly more representative survey made in 2022, where customer service analytics and customer segmentation are respectively in third and fourth place among all use cases.[59] Overall, the findings of the survey show that AI-based language technologies are being used in various European businesses, and their development is expected to expand. The **multilingualism** is striking; only 10% of the companies that use AI-based solutions operate in only one language, while the majority are between one and five languages, and the cases of operating in more than five languages amount to 25%.

[59] https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

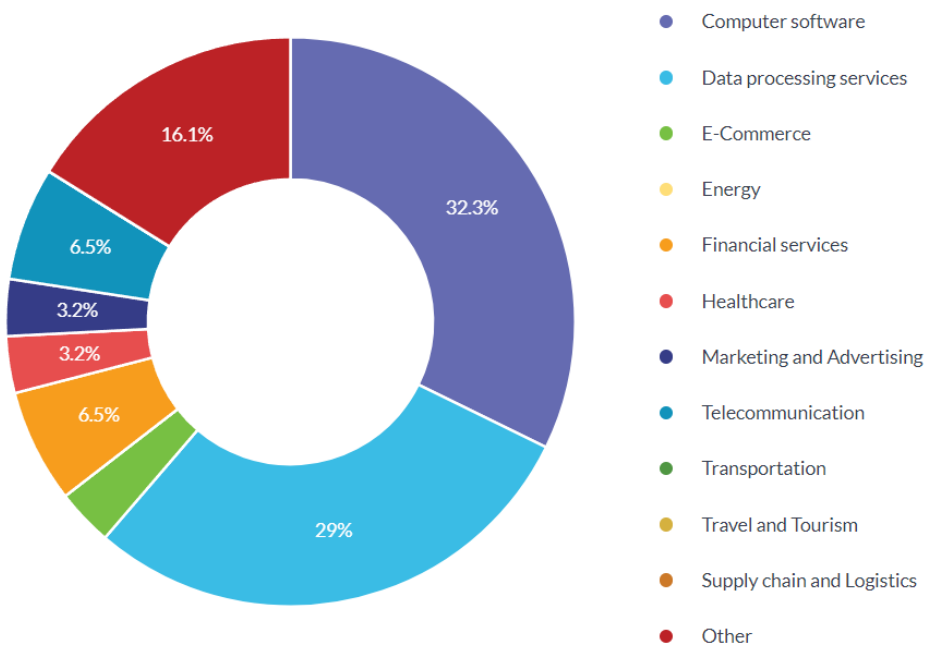

Figure 2: Survey: Which industries are your main areas of operation?

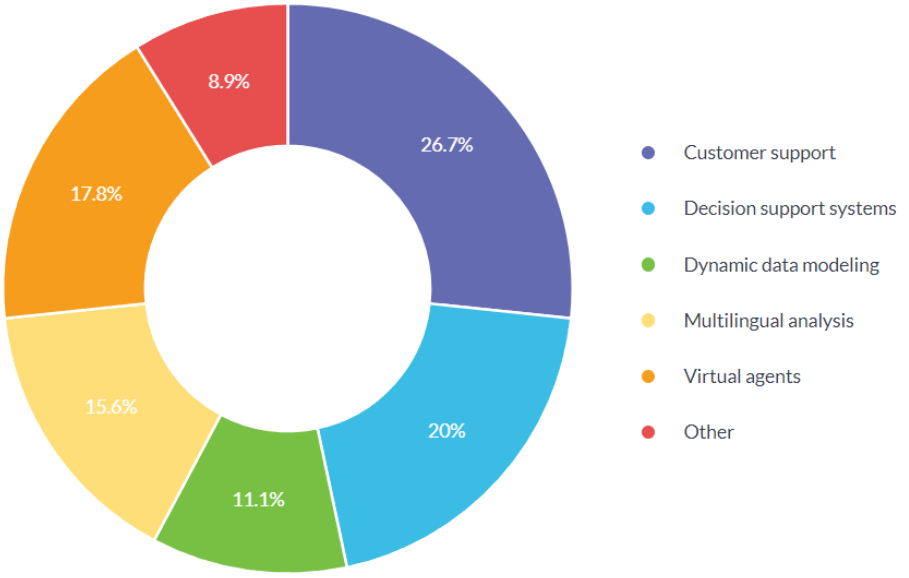

Figure 3: Survey: Does your company use solutions based on Language technologies?

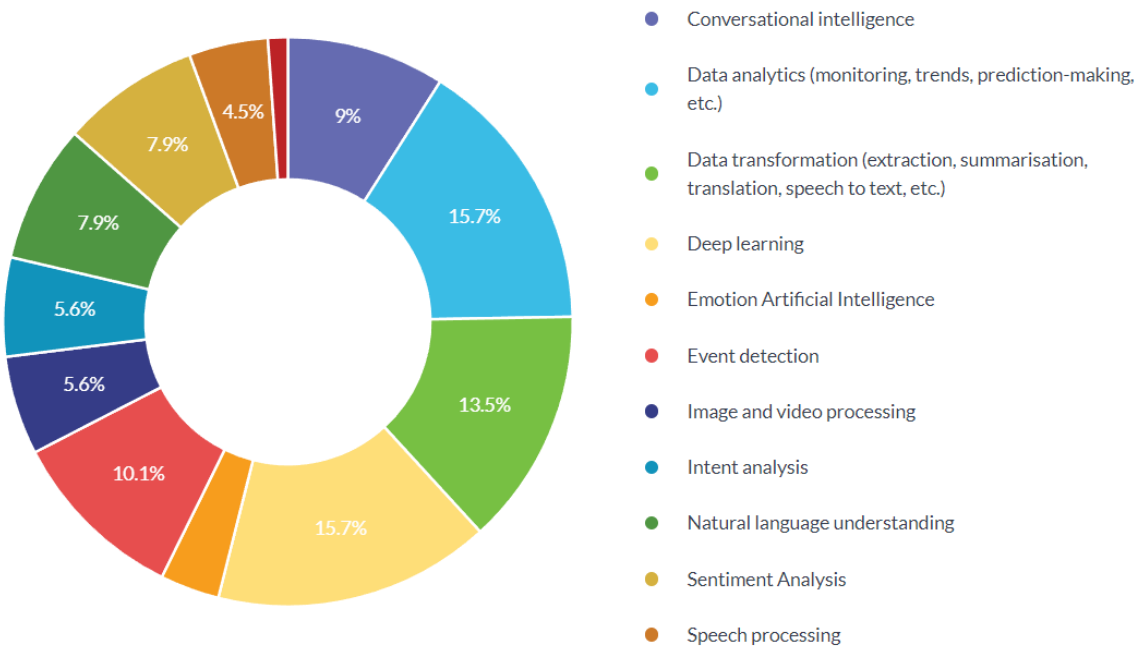

Figure 4: Survey: Which technologies are already employed?

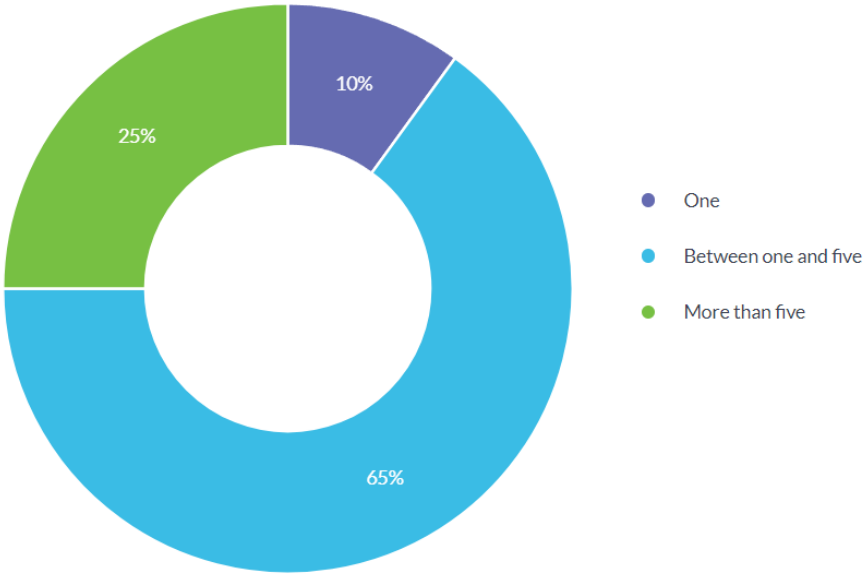

Figure 5: Survey: How many languages are supported?

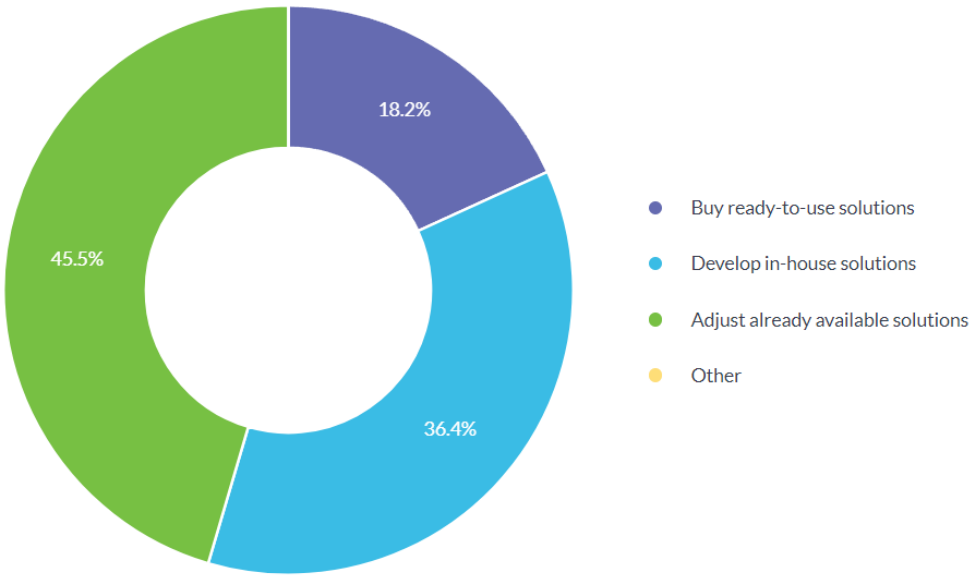

Figure 6: Survey: What do you prefer when integrating new technologies?

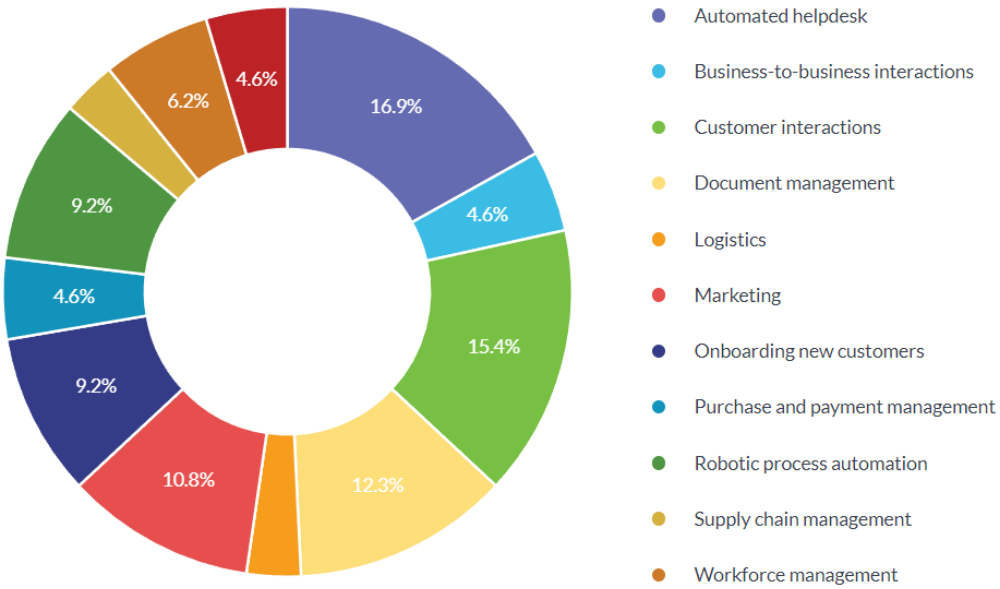

Figure 7: Survey: Which areas of your business can further benefit from AI development?