Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

DeepMinor: Language Models for Multilingual and Multidomain Text Processing in Low Resource Scenarios

Proyecto de Investigación

Rodrigo Agerri Gascón

informatika fakultatea

facultad de informática

Presentado para optar a plaza de Personal Doctor Investigador Permanente con código IDPTCL1-D00141-1 en el Área de conocimiento de Lenguajes y Sistemas Informáticos (Resolución de la Universidad del País Vasco del 18 de diciembre de 2023; BOPV nº 246 del 28 de diciembre de 2023).

Donostia-San Sebastián, 22 de abril de 2024

Summary

Being language the most efficient system for exchanging information, Natural Language Processing (NLP) is one of the most important Artificial Intelligence (AI) based technologies of the current digital transformation. Understanding language is crucial for the success of text analytics and information access applications which depend on the quality of the underlying text-processing tools.

AI-based Large language models (LLMs) have proven their immense potential repeatedly since their introduction several years ago. Most recently, ChatGPT, released by the company OpenAI in late 2022, has demonstrated the extreme disruptiveness of this paradigm-shifting AI technology, which was further improved in early 2023 by the development of GPT-4. These models have been followed by many others, including PaLM by Google, Ernie by Baidu or LlaMA by Meta.

Thanks to these recent advancements, the NLP research field is engaged in a paradigm shift focused on the production and exploitation of these LLMs. In fact, results are improving so much that systems are claiming to obtain human-level performance in laboratory benchmarks when tested on some difficult language understanding tasks.

While impressive, these LLMs have been developed mostly for English[1], they are not public, and have been evaluated almost exclusively on English-centric Natural Language Processing (NLP) benchmarks. These benchmarks are crucial to understand the limitations and possibilities in using these LLMs to improve the state-of-the-art in NLP. Thus, for the large majority of languages and domains, the performance of such LLMs is unknown or it simply cannot be objectively measured. This is due to the fact that either they have not been pre-trained for languages such as Basque or Spanish or because of the lack of readily available benchmarks which would allow to evaluate the Natural Language Understanding and Generation capabilities for those languages. An additional issue of LLMs is that they are still hindered by outdated knowledge and are prone to generate plausible looking content that is actually factually incorrect (known as *hallucinations).*

This project aims to investigate and develop enabling techniques and methods to develop and adapt monolingual and multilingual LLMs to new languages, text genres and domains. In particular, this project will focus on adapting and developing models specially tailored for Basque and Spanish (in addition to English), both for discriminative and generative tasks. We will also work towards filling the current gap on language models in these languages for specific application tasks related with health domain and the fight against misinformation, for which little or no manually annotated data is available[2].

Our progress will be measured by developing new understanding and generation natural language benchmarks and tasks for at least Basque, Spanish and English, focusing on the truthfulness and reliability of the output generated by the LLMs. Thus, we will provide new benchmarks for popular tasks based on text generation and understanding such as Long Answer Question Answering, Explanatory Argument Generation and Inferential tasks for which annotated data for evaluation exists only for English. By doing so we are aiming at significantly improving the state-of-the-art of AI-based Large Language Models in low resource scenarios for languages such as Basque and Spanish thereby contributing to the improvement of Language Technology Applications and its deployment in the current digital transformation.

1 TruthfulQA: Measuring How Models Mimic Human Falsehoods (Lin et al. ACL 2022).
2 Large Language Models Encode Medical Knowledge (Singhal et al., Nature 2023).

Expected Scientific Impact

The Principal Investigator and the research group of which he is member have a strong track record of publishing at national and international level and they will continue work in disseminating results (both research and application related) throughout the duration of the project. This will include the publication of top-ranking journal articles and conference proceedings as well as presentation of the project results at scientific events, shared evaluation tasks, workshops and conferences.

By incorporating the latest insights in AI-based Language Technology, such as large pre-trained language models (LLMs), transfer learning, few-shot and zero-shot capabilities, DeepMinor will leverage and generate carefully designed benchmarks and datasets to advance the state of the art in NLP for English, Spanish, and Basque in several domains and digital sectors. In fact, DeepMinor has the potential to help de-fragment and impact NLP technology on these languages, domains and sectors thereby providing easier access to such technology. For instance, DeepMinor will contribute to information extraction and enrichment of texts by generating Explainable Argumentation and Long Form Question Answering for medical and fight against misinformation applications.

By doing so DeepMinor will also promote multidisciplinary research not only among AI researchers working on NLP, but also with domain-experts from journalism, medicine and communication and citizen digital literacy researchers. This would allow us to also evaluate and investigate the effect of automatically generated explanations for domain-experts and the impact of various strategies of counter-argumentation and its relation with citizen digital literacy and user-awareness. Furthermore, the project will provide new benchmarks for evaluation of explanatory argumentation, truthfulness, Long Form QA generation and inference for at least for Basque and Spanish, addressing a glaring gap on the evaluation of LLMs for these languages. Every generated resource will be publicly distributed under open licences to facilitate more research on this topic and guarantee reproducibility of the published results.

The impact of the project in the academic and industrial communities will be higher due to the resulting technology and linguistic resources: the produced evaluation benchmarks will be very useful not only to researchers in AI and NLP, but also will make possible for the industry to develop information access applications currently infeasible. The produced new software will be distributed under open source licenses, enabling the universal access to a new cutting-edge technology in NLP. The feasibility of the socio-economic impact is boosted by the socio-economic and scientific impact that linguistic tools and resources already contributed by the PI and his group's previous work convey both at national and international levels. Examples of this noticeable impact include thousands of downloads of resources (in which the PI was involved) such as the Multilingual Central Repository (MCR), the linguistic processors such as IXA pipes or our language models and data resources uploaded into the group's Hugging Face repository[3].

3 https://huggingface.co/HiTZ

Table of Contents

1. Novelty and Justification of the Proposal – State of the Art....6
1.1. Background – State of Art....7
1.1.1. Argumentation to Fight against Misinformation....10
1.1.2. Argumentation-based Benchmarking of LLMs in Medical QA....11
1.1.3. Evaluation of Applications based on Argument Generation....13
2. Hypothesis and Previous Contributions of the Team....14
3. Objectives, Methodology and Work Plan....17
3.1. General and Specific Objectives....17
3.2. Methodology....18
3.3. Material, equipment and human resources available for the project....22
3.4. Chronogram and Work Plan....22
3.5. Risks and Contingency Plans....23
4. Justification of the Requested Budget....23
4.1. Qualitative justification of the requested budget for the project....23
4.2. Upgrading of facilities and equipment....24
5. Impact of Results....24
5.1. Expected impact on the generation of scientific and technical knowledge....24
5.2. Social and economic impact outcomes....24
5.3. Expected impact of the proposed activities....25
5.4. Plan for scientific communication and internationalization of the results....25
5.5. Dissemination plan of the results to society....25
5.6. Summary of the data management plan....26
5.7. Results transfer and valorization plan....26
6. Scientific Technical and Training Context....26

1. Novelty and Justification of the Proposal – State of the Art

This proposal targets thematic priority 4. "Mundo digital, industria, espacio y defensa" described in the "Plan Estatal de Investigación Científica, Técnica y de Innovación 2021-2023" (PEICTI) and in particular within axis (1) Digital Transformation and Artificial Intelligence and the among the national R+D+I Artificial Intelligence and Robotics strategic line of "Estrategia Española de Ciencia, Tecnología e Innovación 2021-2027" (EECTI): "Tecnologías del lenguaje; Comprensión profunda del significado del lenguaje".

In recent years, the Natural Language Processing (NLP) community is contributing to the emergence of powerful new deep learning techniques and tools that are revolutionizing the approach to Language Technology (LT) tasks. We are moving from a methodology in which a pipeline of multiple modules was the typical way to implement NLP solutions, to architectures based on complex neural networks trained with vast amounts of text data. This rapid progress in NLP has been possible because of the confluence of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual textual data), 3) increase in High Performance Computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning and transfer learning approaches using Transformers (Devlin et al. 2019; Liu et al. 2020; Torfi et al. 2020; Wolf et al. 2020).

Since the introduction of the Transformer architecture (Vaswani et al. 2017), dramatic progress has been made across a wide range of tasks in AI and NLP. At first, most of the progress was driven by the fact that Transformers were easy to parallelise, scale and be fine-tuned or adapted to a wide range of downstream tasks. At the beginning, most of the progress was driven by simply scaling both the model and the size of its pre-training dataset (Radford et al., 2019), so the capabilities of Transformer-based models were only limited by both the available compute infrastructure and the amount of raw data available for pre-training. However, now, almost six years after the introduction of the Transformer, the hardware and infrastructure to train LLMs is more widely available and we have a much better and also deeper understanding of how far we can scale these models and how much raw data is at our disposal for pre-training (Hoffman et al. 2022). Attention has, thus shifted from simply scaling model sizes more into fine-tuning existing models to follow human instructions, as shown by InstructGPT (Ouyang et al., 2022), rendering these models conversational and more recently, using human feedback and reinforcement learning to better align these models with human intent and to reduce harmful output.

Thanks to these recent advancements, the NLP community is currently engaged in a paradigm shift with the production and exploitation of large, pre-trained transformer-based language models (Han et al. 2021; Min et al. 2021a). As a result, many in the industry have started deploying large pre-trained neural language models in production. For instance, Google and Microsoft have integrated them in their search engines, their flagship product. Compared to previous work, results are improving so much that systems are claiming to obtain human-level performance in laboratory benchmarks when testing on some difficult language understanding tasks.

Furthermore, recent work has shown that pre-trained language models can robustly perform for NLP tasks in a few-shot or even in zero-shot fashion when given an adequate task description in its natural language prompt (Brown et al. 2020; Ding et al. 2021). Surprisingly, fine-tuning pre-trained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Wei et al. 2021; Sanh et al. 2021; Min et al. 2021b; Ye et al. 2021; Aghajanyan et al 2021;

Aribandi et al. 2021).

Despite their impressive capabilities, large Large Language Models (LLMs) do come with severe drawbacks. Currently we have no clear understanding of how they work, when they fail, and what emergent properties they may present, or which novel ways of exploiting these models can help to improve state-of-the-art in NLP. As argued by Bender et al. (2021), it is important to understand the limitations of large pre-trained language models, which they call "stochastic parrots". Some authors call these models "foundation models" to underscore their critically central yet incomplete character (Bommasani et al. 2021). To tackle these questions, much critical multidisciplinary collaboration and research is needed.

Furthermore, these LLMs have been developed mostly for English, they are not public, and have been evaluated almost exclusively on English-centric Natural Language Processing (NLP) benchmarks (Lin et al. 2022).. These benchmarks are crucial to understand the limitations and possibilities in using these LLMs to improve the state-of-the-art in NLP. Thus, for the large majority of languages and domains, the performance of such LLMs is unknown or it simply cannot be objectively measured. This is due to the fact that either they have not been pre-trained for languages such as Basque or Spanish or because of the lack of readily available benchmarks which would allow to evaluate the Natural Language Understanding and Generation capabilities for those languages. An additional issue of LLMs is that they are still hindered by outdated knowledge and are prone to generate plausible looking content that is actually factually incorrect (known as *hallucinations).*

1.1. Background – State of Art

Most NLP systems today are powered by Machine Learning (ML) where predictive models are trained on known data and used to make predictions on new data. The rise of machine learning within AI and NLP started in the 1990s where rather than specifying *how* to solve a task, a learning algorithm induced a model based on a set of *features* representing in the best possible way the training data examples. Thus, complex NLP tasks still require a manually-driven *feature engineering* process to characterize raw data into task useful representations. A few years ago, *Deep Learning* (Lecun et al. 2015) started gaining traction in NLP thanks to mature deep neural network technology, much larger datasets, more computational capacity (notably, the availability of GPUs), and application of simple but effective self-learning objectives (Goodfellow et al. 2016). One of the advantages of these neural language models is their ability to alleviate the *feature engineering* problem by using low-dimensional and dense vectors (aka. *distributed representations*) to implicitly represent the language examples (Collobert et al. 2011). Very recently, the field of NLP faced another relevant disruption with BERT (Devlin et al. 2019). Since then BERT has become a ubiquitous baseline in NLP experiments and inspired a large number of studies and improvements (Rogers et al. 2020). This pre-trained language model recipe has been replicated across many languages. For instance, Basque (Agerri et al. 2020), Spanish (Canete et al. 2020), Catalan (Armengol-Estapé et al. 2021) or Galician (Vilares et al. 2021).

Currently, the NLP field is undergoing a paradigm shift with the rise of ***neural language models*** (also known as Pre-trained Language Models) that are trained on broad data at scale and are adaptable to a wide range of monolingual and multilingual downstream tasks (Han et al. 2021; Min et al. 2021a). Though these models are based on standard *self-supervised* deep learning and *transfer learning*, their scale results in new emergent and surprising capabilities.

In ***self-supervised learning***, the language model is derived automatically from large volumes of unannotated language data. There has been considerable progress in *self-supervised learning* since *word embeddings* (Turian et al. 2010; Mikolov et al. 2013; Pennington et al. 2014; Mikolov et al. 2018) associated word

vectors with context-independent vectors. Shortly thereafter, self-supervised learning based on autoregressive language modelling (predict the next word given the previous words) (Dai and Le 2015) became popular. This approach produced language models such as GPT (Radford et al. 2018), ELMo (Peters et al. 2018) and ULMFiT (Howard et al. 2018). The next wave of developments in self-supervised learning — BERT (Devlin et al. 2019), GPT-2 (Radford et al. 2019), RoBERTa (Liu et al. 2019), T5 (Raffel et al. 2020), BART (Lewis et al. 2020) — quickly followed, embracing the Transformer architecture (Vaswani et al. 2017), incorporating more powerful deep bidirectional encoders of sentences, and scaling up to larger models and datasets.

The idea of ***transfer learning*** is to take the "knowledge" learned from one task (e.g., predict the next word given the previous words) and apply it to another task (e.g., summarization). With transfer learning, instead of starting the learning process from scratch, you start from patterns that have been learned when solving a different problem. This way you leverage previous learning and avoid starting from scratch. Within deep learning, pre-training is the dominant approach to *transfer learning*: the objective is to *pre-train* a deep transformer model on large amounts of data and then reuse this pre-trained language model by *fine-tuning* it on small amounts of (usually annotated) task-specific data. Thus, transfer learning formalizes a two-phase learning framework: a pre-training phase to capture knowledge from one or more source tasks, and a fine-tuning stage to transfer the captured knowledge to many target tasks.

Recent work has shown that pre-trained language models can robustly perform classification tasks in a few-shot or even in zero-shot fashion, when given an adequate task description in its natural language prompt (Brown et al. 2020). Unlike traditional supervised learning, which trains a model to take in an input and predict an output, ***prompt-based learning*** is based on exploiting pre-trained language models to solve a task using text directly (Liu et al. 2021). To use these models to perform prediction tasks, the original input is modified using a template into a textual string prompt that has some missing slots, and then the language model is used to probabilistically fill the missing information to obtain a final string, from which the final output for the task can be derived. This framework looks very promising for a number of reasons: it allows the language model to be pre-trained on massive amounts of raw text, and by defining a new prompting function the model is able to perform ***few-shot*** or even ***zero-shot*** learning, adapting to new scenarios, languages and domains with few or no labeled data. Thus, some NLP tasks can be solved in a fully unsupervised fashion by providing a pre-trained language model with *task descriptions* in natural language (Raffel et al. 2020; Schick and Schutze 2021). Surprisingly, fine-tuning pre-trained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Wei et al. 2021; Sanh et al. 2021; Min et al. 2021b; Ye et al. 2021; Aghajanyan et al 2021; Aribandi et al. 2021).

Multilingual Language Models (MLLMs) such as mBERT (Devlin et al. 2019), XLM-R (Conneau et al. 2020), mT5 (Xue et al. 2021), mBART (Liu et al. 2020), etc. have emerged as a viable option for bringing the power of pre-training to a large number of languages. For example, mBERT (Devlin et al. 2019) is pre-trained with the Multilingual Masked Language Modeling (MMLM) task using non-parallel multilingual Wikipedia corpora in 104 languages. mBERT has the ability to generalize cross-lingual knowledge in zero-shot scenarios. This indicates that even with the same structure of BERT, using multilingual data can enable the model to learn cross-lingual representations. A MLLM is pre-trained using large amounts of unlabeled data from multiple languages with the hope that low-resource languages may benefit from high-resource languages due to a shared vocabulary and latent language properties. The surprisingly good performance of MLLMs in crosslingual transfer as well as bilingual tasks motivates the hypothesis that MLLMs are learning universal patterns (Doddapaneni et al. 2021). Thus, one of the main motivations of training MLLMs is to

enable transfer from high.resource languages to low-resource languages. Thus, of particular interest is the ability of MLLMs to facilitate zero-shot crosslingual transfer from a resource-rich language to a resource-deprived language which does not have any task-specific training data, or to fine-tune more robust language models by using annotated training data in multiple languages.

Text generation, which is often formally referred as Natural Language Generation (NLG), has become one of the most important yet challenging tasks in NLP (Gehrmann et al., 2021). With the recent resurgence of deep learning, various works have been proposed to solve text generation tasks based on different neural architectures (Li et al., 2021b). One of the advantages of these Large Language Models (LLMs) is that they enable end-to-end learning of semantic mappings from input to output in text generation. Transformer encoder-decoder models such T5 (Raffel et al., 2020), BART (Lewis et al., 2020) or a single Transformer decoder block such as GPT (Brown et al., 2020), LlaMA (Touvron et al. 2023) or Mistral (Jiang et al. 2023) are currently some of the standard LLMs for generating high quality text.

As important as it is to develop new rule-based, machine-based or deep learning systems to solve different NLP tasks, it is equally essential to measure the performance of these systems. The most common method to do so is through the use of ***benchmarks***, i.e., according to manually annotated datasets. Leaderboards such as NLP-progress,[4] Allen Institute of AI leaderboard,[5] Papers with code,[6] or Kaggle[7] are meant to encourage participation and facilitate evaluation across many different NLP tasks and datasets. However, most of these evaluation datasets and benchmarks have been developed for English only. For instance, at the time of writing the *Papers with Code* platform includes 1044 English datasets but only 70 for Spanish which appears in sixth position and 13 for Basque.

Current NLP technology allows many advanced ***applications*** which have been unthinkable only a few years ago. NLP is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, text editors, text predictors, automatic translation systems, automatic summaries, inclusive technology, etc (Min et al. 2021a). Its rapid development in recent years predicts even more encouraging and also exciting results in the near future (Han et al. 2021). Currently, our society is developing some fears towards the digital world associated with information distrust of what is published given the growing amount of false content. Our project aims at alleviating these problems by developing new methods and advancing the state of the art in machine reading comprehension of language and fighting against misinformation. This project targets two application scenarios namely, Question Answering and Machine Comprehension for Misinformation and Biomedical Text Analysis.

In summary, recent progress in NLP has been driven by advances in both language model architecture and model pre-training. Transformer architectures have facilitated the building of higher-capacity LLMs for a wide variety of tasks. Open-source libraries such as Transformers[8] may open up these advances to a wider NLP community. The library consists of carefully engineered state-of-the art Transformer architectures under a unified API and a curated collection of pre-trained models (Wolf et al. 2020). Unfortunately, the resources necessary to create the best-performing LLMs are found almost exclusively at US and China technology giants. Moreover, this transformative technology poses problems from a research advancement, environmental, and ethical perspective. For example, models such as GPT-3 are private, anglo-centric, and inaccessible to academic organisations (Floridi et al. 2020). There are also worrying shortcomings in the text corpora used to train these models, ranging from a lack of representation of populations, to a predominance

4 http://nlpprogress.com/
5 https://leaderboard.allenai.org/
6 https://paperswithcode.com/area/natural-language-processing
7 https://www.kaggle.com/datasets?tags=13204-NLP
8 https://huggingface.co/

of harmful stereotypes, and to the inclusion of personal information.

Progress beyond the state of the art

This project aims to investigate and develop enabling techniques and methods to develop and adapt monolingual and multilingual LLMs to new languages, text genres and domains. In particular, this project will focus on adapting and developing models specially tailored for Basque and Spanish (in addition to English), both for discriminative and generative tasks. We will also work towards filling the current gap on language models in these languages for specific application tasks related with *health domain and the fight against misinformation*, for which little or no manually annotated data is available (Singhal et al. 2023).

Our progress will be measured by developing new understanding and generation natural language benchmarks and tasks for at least Basque, Spanish and English, focusing on the truthfulness and reliability of the output generated by the LLMs. Thus, we will provide new benchmarks for popular tasks based on text generation and understanding such as Long Answer Question Answering, Explanatory Argument Generation and Inferential tasks for which annotated data for evaluation exists only for English. By doing so we are aiming at significantly improving the state-of-the-art of AI-based Large Language Models in low resource scenarios for languages such as Basque and Spanish thereby contributing to the improvement of Language Technology Applications and its deployment in the current digital transformation.

In the following two subsections we present the state-of-the-art for the two main domain applications considered: health and fight against misinformation. Furthemore, we also consider the main issues currenly affecting the evaluation of text generation approaches in these two application scenarios.

1.1.1. Argumentation to Fight against Misinformation

Automatic techniques to counteract and mitigate the effects of misinformation are mostly based on explicitly flagging a given message as being suspicious (without any specific explanation to justify the decision). Other approaches include the chatbot service created by the WHO and Facebook to combat misinformation regarding COVID-19[9]. However, the chatbot allows users to get factual and accurate information about the pandemia, it is not a service to counteract misinformation being spread in social media. Therefore, there is a clear lack of AI-based automated approaches to mitigate misinformation by generating appropriate counter-arguments in real time. The closest to this is the work undertaken within the HATEMETER project[10], where they propose using text generation to generate counter-narratives to tackle anti-muslim hate speech. However, the aim of generating counter-narratives is substantially different from generating arguments to address misinformation (Chung et al. 2019, Chung et al. 2021) and it should work under different domain-experts' informed guidelines.

Argument mining aims at extracting natural language arguments and their relations from text (Cabrio and Villata, 2018). Two stages are crucial: (i) argument extraction, understood as the detection of argument components (e.g., claim, premises) and the identification of their textual boundaries; (ii) argument relation extraction: the prediction of the relations (e.g., attack and support) holding between the arguments identified in (i). Most of the work on argument mining has been based on education (Stab and Gurevych 2017) or medical text data (Mayer et al. 2021), and focused on very simple support and attack relations. A number of recent studies have been proposed to address the argumentation synthesis task, which is closely related to the argument generation one. Indeed, these studies propose different approaches to generate claims or reasons

9 https://www.facebook.com/WHO/
10 http://hatemeter.eu/

for a given topic, with a particular stance towards a topic (e.g., (El Baff et al., 2019)). Recently, a number of empirical approaches have been proposed to generate arguments. Park et al., (2019) propose a model called ArgDiver (Argument generation model from Diverse perspectives) which generates multiple sentential arguments that cover diverse perspectives on the given claim. They adopt a Seq2Seq framework and they evaluate their model with the quality of each of the generated sentential arguments, and their diversities. Hua et al., (2019) define a counter-argument generation system to produce paragraph-level arguments with coherent content. Their key feature is to rely on two decoders, i.e., one for text planning — selecting talking points to cover for each sentence to be generated, and a second decoder for content realisation — producing a fluent argument to reflect the decisions made by the text planner. The output consists of longer arguments containing richer information. These approaches are limited to the reformulation of arguments mined from Wikipedia and newswire, which is insufficient to generate high-quality interactive explanatory argumentation to counteract misinformation.

Progress beyond the state of the art

This is a quite recent research area, as the topic has been marginally touched for several years. In addition, DeepMinor will advance the state of the art in argument mining and generation through the capacity to automatically assess the quality of an argument. In case of low quality arguments proposed by users, the argumentation component will be able to react through critical questions (when the argument is not sufficiently elaborated), and requests for additional elaboration (when the argument lacks supporting evidence). The goal is to enhance the critical analysis capabilities of the users with respect to misinformation.

DeepMinor will provide novel AI technology by leveraging the latest advances in NLG to automatically generate domain-expert guided counter-arguments in real time with the aim of counteracting the spread of misinformation that go beyond the support/attack paradigm. This endeavour requires multidisciplinary work between domain-experts on misinformation (fact-checkers, journalists, policy makers, etc.) and AI researchers to generate arguments that fulfil a number of task-specific objectives related to fact-checking and reason-checking (Visser et al. 2020). In this sense, legitimate objectives could be to provide arguments based on factual, rhetoric (assessing the quality of premises and reasoning in persuasive or explanatory texts), by alerting other users of the social media that a particular message might be spreading misinformation (and arguing the justification to do so), or by generating Critical Questions (CQs) to lay the blind spots of the argument provided by a given claim.

1.1.2. Argumentation-based Benchmarking of LLMs in Medical QA

We are currently seeing a dramatic increase in research on how to apply Artificial Intelligence (AI) to the medical domain with the aim of generating decision support tools to assist medical experts in their everyday activities. This has been further motivated by rather strong claims about LLMs in medical Question Answering (QA) tasks, such as that they obtain passing marks for medical licensing exams like the United States Medical Licensing Examination (USMLE) (Singhal et al. 2023, Nori et al. 2023).

Assisting medical experts by answering their medical questions is a natural way of articulating human-AI interaction as it is usually considered that Medical QA involves processing, acquiring and summarizing relevant information and knowledge and then reasoning about how to apply the available knowledge to the current context given by a clinical case. For example, a resident medical doctor preparing for the licensing exams may want to know what and why is the correct treatment or diagnosis in the context of a clinical case (Safranek et al. 2023, Goenaga et al. 2023). This means that a LLM should be able to automatically identify,

access and correctly apply the relevant medical knowledge, and that it will be capable of elucidating between the variety of symptoms, each of which may be indicative of multiple diseases. Finally, it is also assumed that the model will interact with the resident medical doctor in a natural manner, ideally using natural language. Therefore, developing the required AI technology to help, for example, resident medical doctors to prepare their licensing exams remains a far from trivial endeavour.

Nonetheless, and as a crucial first step to address this challenge, the AI ecosystem has seen an explosion of LLMs (both general purpose and specific to the medical domain) reporting high accuracy results on Medical QA tasks thereby demonstrating that LLMs are somewhat capable of encoding clinical knowledge (Singhal et al. 2023). State-of-the-art models include publicly available ones such as LlaMA (Touvron et al. 2023) and the medical-specific PMC-LLaMA (Wu et al. 2023), Mistral (Jiang et al. 2023) and its medical version BioMistral (Labrak et al. 2024), and proprietary models such as MedPaLM (Singhal et al. 2023) and GPT-4 (Nori et al. 2023), among many others.

However, while their published high-accuracy scores on Medical QA may seem impressive, these LLMs still present a number of shortcomings. First, LLMs usually generate factually inaccurate answers that seem plausible enough for non-medical experts (known as hallucinations) (Xie et al. 2023, Xiong et al. 2024). Second, their knowledge might be outdated as the pre-training data used to train the LLMs may not include the latest available medical knowledge. Third, the Medical QA benchmarks on which they are evaluated do not include gold reference argumentative explanations generated by medical doctors providing the required reasoning to support the model's predictions. Finally, and to the best of our knowledge, evaluations have only been done for English, which makes it impossible to know how well these LLMs fare for other languages.

Retrieval Augmented Generation (RAG) techniques have been specifically proposed to address the first two issues, namely, the lack of up-to-date medical knowledge and the tendency of these models to hallucinate (Xiong et al. 2024). Their MedRAG approach obtains clear zero-shot improvements for two of the five datasets on their MIRAGE benchmark, while for the rest the obtained gains are rather modest. Still, MedRAG proves to be an effective technique to improve Medical QA by incorporating external medical knowledge.

Progress beyond the state of the art

In this project we will work to setup multilingual benchmarks for Medical QA. Furthermore, and unlike previous work, our new benchmark will also includes gold reference explanations to justify why the correct answer is correct and also to explain why the rest of the options are incorrect. Written by medical doctors, these high-quality explanations will help to assess the model's decisions based on complex medical reasoning.

We have identified a free data source, CasiMedicos, which consists of Resident Medical Exams or *Médico Interno Residente* (MIR) in Spanish, an exam similar to other licensing examinations such as USMLE, to setup our benchmark. In addition to a short clinical case, a question and the multiple-choice options, CasiMedicos includes gold reference argumentative explanations regarding both the correct and incorrect options.

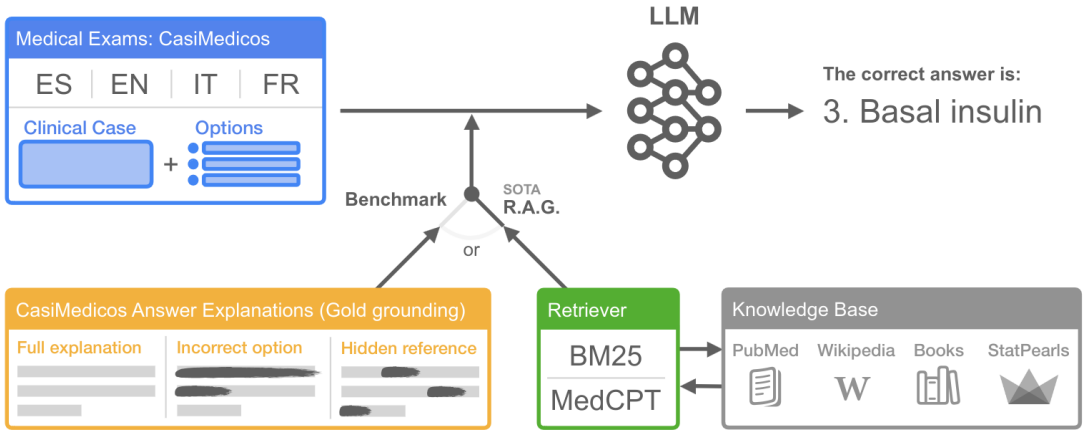

Figure 1: Multilingual Benchmark for LLMs in Medical QA.

Figure provides an overview of our envisaged benchmark. Taking CasiMedicos as the data source, the basic input, without any additional knowledge, to the LLM consists of a clinical case and the multiple-choice options. Furthermore, the model can also be provided with three types of gold reference explanations (or gold knowledge grounding) extracted from the CasiMedicos explanations; (i) the full gold explanation as written by the medical doctors; (ii) only the explanations regarding the incorrect answers and, (iii) the full gold explanation with explicit references to the possible answers hidden. Finally, we can also apply automatic knowledge retrieval approaches such as MedRAG to provide LLMs with automatically obtained up-to-date medical knowledge. Thus, in this benchmark it is possible to compare not only whether the MedRAG methods improve over the basic input with no external knowledge added, but also to establish the differences in performance of LLMs (with or without RAG) with respect to results obtained when gold reference explanations are available. An additional benefit of it being multilingual is that we get to compare LLMs performance not only for English, but also on popular languages such as Spanish.

1.1.3. Evaluation of Applications based on Argument Generation

NLG tasks such as explanatory argumentation generation to counteract misinformation as proposed in DeepMinor presents a considerable evaluation challenge. Thus, while it is possible to use usual distance-based metrics to evaluate the text generation (Gehrmann et al., 2021) such as ROUGE, BLEU or Bertscore (Zhang et al. 2019), other works have proposed to use quality-based metrics such as Diversity and Novelty to evaluate the capacity of the model to generate diverse/varied responses and the ability to generate sequences different from the data seeing during training/fine-tuning (Wang and Wan, 2018, Chung et al. 2020). However, a proper evaluation of the explanatory arguments generated in DeepMinor to counteract misinformation requires to consider task-specific issues not taken into account in previous NLG or argumentation work. This implies to evaluate the quality of the generated counter arguments regarding the supporting evidence found in trusted resources.

Progress beyond the state of the art

DeepMinor will advance the state of the art on evaluation of generation of argument-based explanations by considering argument-related (argument quality assessments, their role in persuasive discourse, level of elaboration, etc.) as well as task-specific criteria (is the response or explanation generated on-topic, namely, does it achieve address the main issue raised by the misinformation item, is the choice of argument type the most appropriate (factual, rhetorical, providing a simple alert) and so on. The second type of criteria

specifically requires the involvement of domain-experts given that social science research on best strategies to answer perceived misinformation is still not conclusive. We will need to evaluate the resulting counter-arguments with respect to the targeted audience (e.g., young and elderly people). The notion of "good" counter-argument is not uniform but must coincide with the receptivity of the interlocutor, more sensitive to certain modes of explanations and indifferent to others. This means that, given a certain topic and the identified form of disinformation, different nuances can be privileged in the generated counter-argument(s). Studying the correlation between quantitative, automatic metrics and qualitative criteria to evaluate NLG of misinformation counter-arguments will also be a novel contribution of DeepMinor.

2. Hypothesis and Previous Contributions of the Team

The NLP community is currently engaged in a paradigm shift with the production and exploitation of large, pre-trained transformer-based language models (Han et al. 2021; Min et al. 2021a). This paradigm shift means that we have only just started to scratch the surface of the new possibilities offered by these large pre-trained language models. **Our initial hypothesis in DeepMinor** is that we can obtain substantial improvements in many NLP tasks by (i) generating and exploiting new language models for Basque, Spanish and English by taking into account a multitask objective during the pre-training; (ii) exploring novel ways, such as prompting, of exploiting these language models to improve NLP results on zero-shot and few-shot settings (without or very little training data for the target language or task at hand); (iii) by addressing language understanding tasks by text generation; (iv) by leveraging pre-trained language models and knowledge bases, (v) developing new benchmarks and datasets for evaluating and assessing our progress towards Natural Language Understanding; (vi) to apply the newly developed techniques to improve the state-of-the art in language understanding, especially for settings with few or non-existing training data and (vii) by developing a number of advanced content-based domain applications for the main official languages in Spain (or at least Basque and Spanish) plus English in the two main application scenarios considered, namely, the medical domain and the fight against misinformation.

In the DeepReading project (RTI2018-096846-B-C21 MCIU/AEI/FEDER, UE) we obtained state-of-the-art results in cross-lingual and multilingual NLP tasks by researching new deep learning methods based on neural networks and language models. Previous contributions relevant to DeepMinor include the following:

Deep Multilingual Text Processing: Pre-trained language models are the core building block in current deep learning based NLP. The team has experience building language models for several languages, including Spanish and Basque, as well as multilingual models. For Basque, we built BERTeus[11] (Agerri et al. 2020), a model following the BERT architecture, as well as models following the RoBERTa architecture[12]. For Spanish, we built two models based on RoBERTa (IXABERTes v1[13] and v2[14]). Finally, we also built a multilingual model for Basque, Spanish and English, following the multi-BERT architecture.[15]

Novel paradigms for the exploitation of language models: Regarding prompting and zero-shot learning, the team successfully combined prompting templates with label verbalization, reformulating event extraction tasks as entailment problems (Sainz et al. 2021a), showing the zero-shot learning capabilities of language models. The team has also contributed to an exhaustive survey of recent work that uses large language

11 https://huggingface.co/ixa-ehu/berteus-base-cased
12 http://www.ixa.eus/euscrawl/#models
13 http://www.deeptext.eus/resources/ixabertes-v1.zip
14 http://www.deeptext.eus/resources/ixabertes-v2.zip
15 https://huggingface.co/ixa-ehu/ixambert-base-cased

models to solve NLP tasks via pre-training then fine-tuning, prompting, or text generation approaches (Min et al. 2021a).

Summarizing, the research group in DeepMinor is a all well-known international player in Language Technology, and specially on broad coverage Natural Language Processing for Basque, Spanish as well as English. Thus, our research groups have been jointly involved in the construction and enrichment of NLP tools and semantic resources within several national (ITEM, HERMES, SENSEM, KNOW, TEXT-MESS, KNOW2, SkaTer, TUNER, DeepReading, VIGICOViD) and European research projects (ACQUILEX, ACQUILEX-II, EuroWordNet, MEANING, KYOTO, PATHS, OpeNER, NewsReader, READERS, LIHLITH, BETTER).

Furthermore, regarding Argumentation approaches to NLP applications, DeepMinor will benefit from collaborations with other international project on which the partners are involved such as Antidote: Argumentation-driven Explainable Artificial Intelligence for Digital Medicine. CHIST-ERA (European Comission), INT-Acciones de Programación Conjunta Internacional PCI2020-120717-2 where the objective is to study argumentation-based explainability techniques in the medical domain.

Applications:

The central concept in DeepMinor linking argumentation with *health applications* and *misbehaviour detection and mitigation* is that of explainability, or explainable AI (XAI). Thus, interactively generated explanatory arguments facilitate trustworthiness between human (domain-experts) and AI technology, according to the EU Ethics Guidelines for Trustworthy AI[16]. For a socially and ethically delicate issues such as those related with health or misinformation, argumentation-based XAI would allow to comply with requirements such as transparency, explainability, communication, societal friendliness and accountability, because it would allow human experts to better understand the machine predictions. According to Dufour (2017), human agents require explanations when they accept a given conclusion but challenge the premises used to reach it. In these situations, the human's goal is understanding, which is typically reached by some kind of argumentative reasoning.

Regarding misinformation, human fact-checkers publish the results of the fact-checking process explaining why a given social media item (text message, video, image) is spreading misinformation. Recent studies suggest that even short (280 characters) refutations are preferable to no refutation at all. In fact, it has been empirically established that argumentative refutations are more effective than simple false-tag flags, as most social media do, which could be even harmful (Ecker et al. 2020, Ecker et al. 2017). Taking this into account, the aim of AI technology to fight misinformation would have to be of assistance to scale both explainable detection and counter-argument generation to the large amount of misinformation (impossible to address manually) that is continuously being spread within social media.

However, while current AI technology is performing reasonably well for automatic detection (Augenstein et al. 2021), most existing work explaining the predictions of the AI technology is still in its infancy as previous research has focused on highlighting fragments in the input of the message or on generating simple summaries of the evidence used to make the prediction. For mitigation the situation is bleaker, as there is no previous research on possible automatic AI-based strategies to generate automatic responses in social media explaining why a given message is considered to be spreading misinformation.

Considering this, DeepMinor approach will consist of automatically generating explanatory arguments as an interactive type of reasoning which underlies those desirable processes in an AI system to fight misinformation. Those processes are currently being performed by human domain-experts (fact-checkers and

16 https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

other domain-experts which are users of social media), namely, explaining the results of a multimodal fact-check (detection) and generating counterspeech or refutations based on argumentation to aim to reduce misinformation-congruent reasoning in social media.

Furthermore, it will address, in the mitigation phase, the counteracting of misinformation by generating automatic counter-arguments and Critical Questions (CQs) with the aim of mitigating the effects of the spreading of misinformation. Finally, by experimenting on socially/domain-expert guided argumentation generation by few-shot learning, it will help to generate high performing and deployable technology for each of the domains/topics of interest related to misinformation.

This vision is made possible by the huge leaps in performance in NLU and NLG provided by the Transformer-based language models on which DeepMinor will investigate new methods to exploit them in few-shot learning settings. Furthermore, it will also design strategies and benchmarks to evaluate veracity and truthfulness for the languages of interest of the project. Additionally, the project aims to follow recent trends on human-centric AI where humans are by design in the loop. Being aligned with many of the hot topics in AI research (argumentation, deep fakes, multimodality, few-shot learning, citizen digital literacy) DeepMinor will benefit from the advances being achieved on those topics.

Specific bechmarks and experiments will be designed and tailored to the two topics (pseudoscience and inmigration) in collaboration with the best known professional fact-checkers in Spain (Maldita.es, Newtral) part of the DeepMinor's End-User Advisory Board (EUAB) and highly interested in the project's outcomes. Such experiments will look to ensure that the format, content and objectives of the argument-based explanations, for both detection and mitigation, are optimal to efficiently convey information to domain-experts and effective against the spread of misinformation in social media.

In the medical domain, argumentative explanations are pervasive and very few works, if any, have addressed the issue of how to use LLMs and AI to learn to automatically argue about a given diagnosis or treatment. One of the reasons is the lack of available datasets with reference gold explanatory arguments which is why most of the literature has been focusing on Medical QA, namely, on how to answer multiple-choice question tests automatically. We have identified a resource, however, that may help to mitigate the lack of data to learn argumentation-based explanations in the medical domain.

Every year the Spanish Ministry of Health releases the previous year's Resident Medical exams or Médico Interno Residente (MIR) which, includes a clinical case, the multiple choice options and the correct answer. The MIR exams are then commented every year by the CasiMedicos MIR Project 2.0[17] which means that CasiMedicos medical doctors voluntarily write gold reference explanations (full gold explanation providing reasons for both correct and incorrect options.

Such resource could be annotated with argumentative components and relations to perform various novel tasks: (1) setup a new multilingual benchmark for Medical QA, the first of its kind; (2) investigate how to learn to automatically generate argumentative explanations to justify the answers given in the benchmark; (3) to study how to incorporate RAG methods to palliate the lack of readily available knowledge to generate factually correct medical explanations; (4) to address the lack of annotated data by researching on crosslingual transfer and few-shot techniques to perform both Argument Mining and Generation.

17https://www.casimedicos.com/mir-2-0/

3. Objectives, Methodology and Work Plan

In the current context of paradigm shift within the NLP community (Han et al. 2021; Min et al. 2021a), DeepMinor will aim to develop and adapt new language models (i) for specific domains; (ii) and to explore novel methods of exploiting such language models based on the use of prompts and instruction-based fine-tuning for text generation, which we believe will help these pre-trained models to ground their knowledge improving truthfulness reliability, factuality and generalization skills; (iii) which will be evaluated on new benchmarks for Basque, Spanish and English based on tasks such as Explanatory Argument Generation, long form Question Answering and Inference.

3.1. General and Specific Objectives

While currently available models for Basque and Spanish are mostly developed on NLU tasks, DeepMinor will build models that are also capable to deal with text generation tasks, which have shown to generalize better and yield good results work in zero-shot and few-shot scenarios. We will also work towards filling the current gap on language models in these languages for specific domains, such as Health, Education and Social media. Regarding text processing applications, the PI has ample experience developing NLP tools, both basic NLP modules (Agerri et al, 2014, Agerri and Rigau 2016, Agerri et al. 2020, Agerri and Agirre 2023) as well as advanced semantic processing tools in many languages. Taking these issues into account, we address the following specific objectives:

1 To compile large scale datasets and corpora to pre-train and adapt new text generation models for Basque and Spanish (WP2), especially focused on long form QA and inference for explainability and argumentation in both application domains, health and misinformation.

2 To apply domain-specific language models to improve state-of-the-art results on applications related with Long Form Question Answering, Explanatory Argument Generation and Inference (WP3).

3 To improve qualitative and quantitative evaluation of text generation-based tasks by providing new benchmarks for Basque and Spanish focusing on truthfulness and factuality; organize a shared task to motivate work on this topic. Research correlation between qualitative and quantitative distance-based metrics to evaluate the effectiveness of the provided argumentation-based explanations or long form answers (WP4).

4 To come up with novel strategies, such as prompting, to exploit language models for text generation to perform better in zero-shot and few-shot scenarios in cross-lingual settings. This will be crucial to improve results on common tasks but also to mitigate the lack of training data for a given language or specific domain (WP3).

5 To leverage the generated language models to develop state-of-the-art, ready-to-use, linguistic processors for common NLP tasks, such as lemmatization, NER, SRL, POS tagging, among others (WP2).

6 To investigate techniques on the evaluation of text generation tasks defined in the benchmarks above. The idea would be to minimize the amount of manual work to a minimum while maximizing human correlation and task specificity, something that current overlap-based metrics lack (WP4).

3.2. Methodology

As it can be seen in the description of the tasks structured in WPs below, the duration of DeepMinor is 36 months. To achieve the objectives explained earlier, the work has been organized in 3 technical Work Packages, plus a WP1 Management and a WP5 for Dissemination and Exploitation. The technical work is subdivided into three main activities: (i) Data Compilation and LLMs Generation (WP2); (ii) Methods to exploit LLMs for multilingual (Basque, Spanish and English) and few-shot settings in low resource scenarios (WP3), (iii) Benchmarking and Evaluation (WP4), focused on argumentation-based explainability approaches to health applications and misinformation.

Work package number	**2**
Work package title	**Data Compilation and Large Language Models Generation**

Objectives

DeepMinor will adapt and build state-of-the-art multilingual language models for Basque, Spanish as well as English. The models will be based on news technologies, architectures and training paradigms that allow a better generalization between domains and languages. We will build generative models that allow the generation of text in these languages, which is needed in tasks such Long Form Question Answering, Explanatory Argument Generation or Inference. Besides, the project will also build language models adapted to specific domains of Health, Education, Social media.

Task 2.1 Resources compilation.

This task includes identifying data sources and collecting the corpus. The corpora compiled in this task will include general purpose text, as well as text from the specific domains needed in task 2.3. Special care will be put towards building a corpus that is both diverse and inclusive. Among other aspects, the corpus will include dialectal and regional usages, and a proper representation of genres and cultural minorities will be guaranteed.

The corpus will include news corpora, literary corpora, where the job will be focused on collecting books in digital format; Web corpora based on current crawling datasets such as Common Crawl, OSCAR, cc100 or mc4; and domain corpora of the Health and Education domains, as well as from Twitter.

Task 2.2 Multilingual models for Basque, Spanish and English.

While language models for the languages covered in the project exist, they are mostly focused towards NLU applications. In this task we will complement this development with new models based on state-of-the-art architectures that are focused on two main objectives. One is to develop generative models that will allow text generation. The second objective is to use prompt engineering and instruction-based fine-tuning to obtain models that are adapted to zero-shot or few-shot scenarios, and are able to leverage training data from resourceful languages (e.g. English) to languages with less (manually annotated) resources such as Basque and Spanish. As a result, we expect a significant advance on applications that require generating text, as well as those applications for which there is a lack of annotated training data.

Task 2.3 Adapting Language models to specific domains.

Open domain language models, which are trained on corpora such as Wikipedia, News, books, etc,

suffer a significant drop in performance when used to develop applications that deal with text of specific domains, such as medical text or social media. In this task we will build models specifically crafted to work with text from the health, education and social media domains, including tweets. The models will be built using the domain specific corpora collected in Task 2.1. As a result, we expect a significant boost in the performance in such domain-specific applications and the languages covered within the project.

Task 2.4 Language Models for higher-level semantic tasks.

Traditional pipelines, where the outcome of lower-level linguistic interpretation tasks provide the input for higher-level interpretation tasks suffer from error propagation. This task will provide various neural network architectures for higher-level semantic processing tasks (for instance, NER, Entity Linking, Detection and normalization of temporal expressions, Semantic Role Labelling, Coreference, Discourse Relations, Polarity classification, etc.) for the target languages (Basque, English, Spanish). It will explore to what extent higher-level tasks can be learned in an end-to-end manner, staying agnostic of lower level tasks and where cascaded learning using supervision from both lower and higher layers works best.

Deliverables (brief description) and month of delivery

D2.1 First version of DeepMinor LM models and resources (M12)

D2.2 Second version of DeepMinor LM models and resources (M24)

D2.3 Third version of DeepMinor LM models and resources (M36)

Work package number	**3**
Work package title	**Applying Language Models on Multilingual and Low Resource Scenarios**

Objectives

Develop novel ways to exploit the full potential of large language models, including prompting, generation instruction-based fine-tuning. The objective of such exploitation paradigms is two-fold: (i) to improve the overall language understanding capabilities of language models, and (ii) to make them usable for a great variety of applications and languages with minimal preparation effort, through zero-shot and few-shot learning.

Large language models are expensive to train. Given the required resources, fine-tuning a large language model to exploit its potential in a novel task, does not seem to be the best option. To address this issue, in this WP we propose to further explore novel paradigms to train and exploit language models.

T3.1 Research on prompting schemes for language models.

This task will research on diverse aspects of prompting for task adaptation. More concretely, the relation between prompt engineering and design (both for discrete and continuous embeddings) and answer engineering will be analyzed, using different supervised tasks and datasets reformatted consequently. We aim at combining prompt tuning techniques with reinforcement learning strategies to automatically generate the best prompts for different tasks. Instruction-based techniques will also be explored for answer engineering, trying to better align them with the

designed prompts. The objective is to maximize the transferability of the prompts across different tasks.

T3.2 Controlled textual generation.

This task will develop novel ways of prompting and training generative language models to control their text generation abilities. We aim at conditioning generative models not only to generate contents aligned with a specific topic and objective, but also to control the form and style of the generated text. The idea is to extract from the training text the interesting features for a given application, such evidence related to the task (via Retrieval Augmented Generation), the topic and so on. During the pre-training step, the extracted information can be added as explicit learning signals, making the language model learn how the features are related with the actual text. In this way, we will analyze whether the language model can generate structures, forms and/or topics which were not in the training corpus, to check the generalization abilities of the model. The envisaged applications include Long Form Question Answering and Argument Generation, where extracted evidence and features from text can be crucial for domain-specific applications in the health domain or in the generation of counter-arguments to combat misinformation.

T3.3 Zero-shot and few-shot learning.

This task aims at exploring new ways to approach zero-shot and few-shot learning. The main idea is to use proxy tasks for training language models (NLI or QA tasks), in order to take advantage of strong supervised learning signals from existing datasets. At inference time, any kind of task will be automatically formatted as the proxy task, enabling zero-shot and few-shot transfer. Task-specific post-processing steps will be implemented to adapt the proxy task to the target task. More specifically, the relation between prompt engineering and design and answer engineering will be analyzed. We aim at combining prompt tuning techniques with reinforcement learning strategies to automatically generate the best prompts for different tasks. The objective is to maximize the transferability of the prompts across the different tasks in the project (multimodal detection, explainable argumentations and counter-argument generation).

Deliverables (brief description) and month of delivery

D3.1 First results on language models and prompting (M12)

D3.2 Text generation, zero-shot and few-show results (M24)

D3.3 Aspect-controlled text generation (M36).

Work package number	**4**
Work package title	**Evaluation and Assessment**

Objectives

The objective of this work package is to measure the research progress via new benchmarks generated specifically for Basque and Spanish (plus English) for the evaluation of tasks based on text generation (WP3): Long Form Question Answering, Explanatory Argument Generation and Inferential tasks. These benchmarks currently are available only for English and we will address

them in a number of specific domains, including health, misinformation and education.

Task 4.1 Design and development of evaluation benchmarks

The most adequate strategy will be chosen for any development of the evaluation benchmarks: crowdsourcing, re-purpose of existing datasets from other languages (mostly English). The resulting test sets will then be made available to the community and leveraged for evaluation campaigns in forums such as IberLEF, CLEF or SemEval.

Current candidates to start adapting benchmarks for Basque and Spanish include TruthfulQA (multidomain Question Answering benchmark to measure truthfulness of LLMs in English) and MultiMedQA, a multi-task, seven multiple-choice dataset on various types of Question Answering for the medical domain). We will also aim to build a multidomain and multilingual argument generation benchmark which would address both misinformation counterargumentation and medical explanatory argumentation (based on CasiMedicos data). We will test the LLMs developed in WP2 with techniques from WP3 setting new baselines for Basque and Spanish. We will announce the new collection and create a leaderboard.

Task 4.2 Shared Multilingual Task on Measuring Truthfulness of LLMs.

We will propose the shared task at evaluation forums such as IberLEF or CLEF. The data will be based on the benchmarks generated in the previous task. The main task will proposed for both Basque and Spanish. The main objective of this task will be to generate interest in the evaluation of LLMs using modern benchmarks but also to investigate both the qualitative and quantitative evaluation of text generation approaches, which is by no means a solved issue.

Task 4.3 Automatic evaluation of generated counter-arguments and explanations.

The resulting model to generate counter-arguments will be evaluated along with standard metrics, namely, BLEU and BertScore (Zhang et al. 2019) concerning the lexical and semantic generation performances, but also with more recent metrics such as *novelty* and *diversity* (Wan and Wang 2018), which evaluate ability of a model to generate diverse/varied responses that are different from the data seeing during training/fine-tuning. In addition, the explanatory arguments generated in DeepMinor require to consider task-specific parameters, including evaluating the quality of the generated counter arguments regarding the supporting evidence found in trusted resources. Furthermore, correlations between automatic and user-based evaluation of T4.4 will be studied.

Task 4.4 User-based evaluation.

User-based evaluation will have two objectives: (i) to measure the effect of the counter-argumentation generated as counterspeech for misinformation mitigation - reduction of misinformation congruent-reasoning; (ii) to evaluate the resulting counter-arguments with respect to the targeted audience. The notion of "good" counter-argument is not uniform but must coincide with the receptivity of the interlocutor, more sensitive to certain modes of explanations and indifferent to others. This means that, given a certain topic and the identified form of disinformation, different nuances can be privileged in the generated counter-argument(s). The generated argumentation will be evaluated along three axes (Lewinski, 2019): 1) validity of inferences used; 2) quality of interactions generated; 3) epistemic quality of conclusions in the given domain. Both formal and informal approaches to argumentation typically define quality in terms of exhaustion: there are no relevant questions to be asked without repeating previous questions. At this stage, the goal of the

argumentation is satisfied: all *critical questions* are answered and the understanding of the questioner - the ultimate goal of explanation - is reached. Measurable factors such as time spent by each user interacting with the system or the number of repeated questions are taken as indicators of this model. A sample of the generated argumentation will be given to fact-checkers in the EUAB for evaluation, through a questionnaire, of the consistency and of the naturalness of the argumentation.

Deliverables (brief description) and month of delivery

D 4.1: Datasets available (M18)

D 4.2: Shared tasks results (M24)

D 4.3: Final datasets and benchmarks available (M36)

Three milestones, at months 12, 24 and 36 have been settled to assess the intermediate results and represent the end of crucial phases of the project, a first Analysis phase and three Development-Evaluation cycles. In a first step, in WP2 we will define the requirements and collect and generate the required evaluation data for the project. WP3 will primarily focus on the advancement of the state of the art on text generation techniques for Argument Generation. Long Form Question Answering and also on exploiting language models in novel ways, such as prompting, to extract more generalizable and grounded knowledge from the language models. WP4 is based on the techniques resulting from WP3 to establish novel benchmark for the evaluation of LLMs in Basque, Spanish and English.

3.3. Material, equipment and human resources available for the project

The IXA research group and HiTZ Center owns several high-performance servers (including several GPUs servers) with the necessary storage and facilities to accomplish the project objectives. In particular, the group owns 16 general purpose x86-64 multiprocessor GNU/Linux servers with up to 256GB RAM, 4 SPARC Solaris servers used for basic NLP processing, and one HPC cluster comprising 8 x86-64 nodes with 16 cores and 128GB RAM each. Additionally, the group owns servers with 20 NVIDIA small GPUs of 12GB, 1 server with 4 V100 of 32Gb, 1 with 4 A100 of 80Gb, 2 with 8 A100 of 80Gb, and 1 with 2 A30 of 24Gb. While this computing power places our team in a good position to work with large language models, we have asked for an additional 4 NVIDIA A100 GPUs server. This is required to pre-trained large models for text generation, such as T5, in parallel mode. Finally, we own network storage servers offering 25 TB of storage capacity.

3.4. Chronogram and Work Plan

Duration of tasks in 3 month periods (trimester), totalling T12 trimester, Month of Deliverables with D. Four milestones, at months 6, 12, 24 and 36, have been fixed to assess the intermediate results and to identify the end of crucial phases of the project.

Task	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
2.1				D2.1				D2.2				D2.3
2.2								D2.2				D2.3

Task	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
2.3								D2.2				
2.4								D2.2				D2.3
3.1				D3.1				D3.2				D3.3
3.2								D3.2				
3.3								D.3.2				D3.3
4.1				D4.1								
4.2								D4.2				
4.3								D4.2				
4.4												D4.3

3.5. Risks and Contingency Plans

With respect to the possible risks that may arise during the execution of the project and their corresponding contingency plans, we have identified risks and mitigation actions:

Risk	Mitigation action
Difficulties in collecting open licensed source data to construct the large datasets and models n WP2 and publicly released them together with the generated language models.	The partners have a number of ongoing projects with public institutions and media publishers. If negotiations to release data fail, we have a number of publishers identified which release all their publications under creative common licenses.
Generation models too large to be pre-trained in our GPU infrastructure.	While our GPU infrastructure is improving, if needed we have experience in obtaining projects from RES to use their computing power.
Shared tasks not accepted at evaluation forums	Create a website proposing the task, announce it and create a leaderboard where we can compare different approaches

4. Justification of the Requested Budget

4.1. Qualitative justification of the requested budget for the project

DeepMinor requests support for hiring 1 postdoc (postdoc) for the duration of the project and 1 computer scientist for 1 year (no-doc). Among the priority areas of DeepMinor, adapting and developing LLMs (WP2), Exploitation of Language Models in Low Resource Scenarios (WP3) is where our group envision most of the experience to achieve the objectives listed. Furthermore, the second computer scientist is justified by the

high level requirements of new deep learning architectures, frameworks and high-performance systems with multiple CPUs and GPUs. Thus, a new role for the administration, deployment, maintenance and support for the implementation of Deep Learning Language models for NLP is also required for the satisfactory completion of the project. DeepMinor also requests a budget for annotation and post-edition of existing benchmarks for English, such as TruthfulQA or MultiMedQA, in addition to generating our own datasets for Argument Generation.

4.2. Upgrading of facilities and equipment

While the computing power of our team places us in a good position to work with large language models, we have asked for an additional 4 NVIDIA A100 GPUs server. This is required to pre-trained large models for text generation, such as T5 or LlaMA, in parallel mode. A new server is crucial to support all the experimentation planned with LLMs in the project. Furthermore, we ask for budget to install 10GbE switches in the racks of CPD adding high-velocity adapters those servers that do not offer the possibility of using the 10GbE network. The improvements and expansions in the data network infrastructure that interconnects the servers is an important factor in optimizing the possibilities of the high-speed network. This will result in an improvement in the performance of the equipment, which is considered necessary for the initiation or consolidation of the research line proposed by the project.

5. Impact of Results

5.1. Expected impact on the generation of scientific and technical knowledge

The PI and the research group of which he is member have a strong track record of publishing at national and international level and they will continue work in disseminating results (both research and application related) throughout the duration of the project. This will include the publication of top-ranking journal articles and conference proceedings as well as presentation of the project results at scientific events, shared evaluation tasks, workshops and conferences.

By incorporating the latest insights in AI-based Language Technology, such as large pre-trained language models (LLMs), transfer learning, few-shot and zero-shot capabilities, DeepMinor will leverage and generate carefully designed benchmarks and datasets to advance the state of the art in NLP for English, Spanish, and Basque in several domains and digital sectors. In fact, DeepMinor has the potential to help de-fragment and impact NLP technology on these languages, domains and sectors thereby providing easier access to such technology. For instance, DeepMinor will contribute to information extraction and enrichment of medical texts to learn generating Explainable Argumentation and Long Form Question Answering.

5.2. Social and economic impact outcomes

In terms of social and economic impact, DeepMinor will also promote multidisciplinary research not only among AI researchers working on NLP, but also with domain-experts from journalism, medicine and communication and citizen digital literacy researchers. This would allow us to also evaluate and investigate the effect of automatically generated explanations for domain-experts in health applications and the impact of counter-argumentation on social media users and its relation with citizen digital literacy and user-awareness. Furthermore, the project will provide new benchmarks for evaluation of explanatory argumentation, truthfulness, Long Form QA generation and inference for Basque and Spanish, addressing a

glaring gap on the evaluation of large language models for these languages. Every generated resource will be publicly distributed under open licences to facilitate more research on this topic and guarantee reproducibility of the published results.

The impact of the project in the academic and industrial communities will be higher due to the resulting technology and linguistic resources: the produced evaluation benchmarks will be very useful not only to researchers in Artificial Intelligence and NLP, but also will make possible for the industry to develop information access applications currently infeasible. The produced new software will be distributed under open source licenses, enabling the universal access to a new cutting-edge technology in NLP. The feasibility of the socio-economic impact is boosted by the socio-economic and scientific impact that linguistic tools and resources already contributed by PI and his group's previous work convey both at national and international levels. Examples of this noticeable impact include thousands of downloads of resources (in which the PI was involved) such as the Multilingual Central Repository (MCR), the linguistic processors such as IXA pipes or our language models uploaded into Hugging Face repository.

5.3. Expected impact of the proposed activities

LLM-driven chatbots are currently revolutionising information technology as we know it, with substantial implications for and disruptions in society, research, and industry at large. ChatGPT and similar technologies developed by other US or Asian tech giants can be used as writing assistants, personal helpers, general problem solvers, text robots and sparring partners for everyday tasks, challenges, and situations in our personal or professional lives. They can be applied in every single domain, from customer service to healthcare, from mobility to education, in finance, insurance, e-commerce and many others.

However, most of the LLMs have only been tested on English-centric evaluation benchmarks and remain proprietary, which represents a glaring gap for the research and development of LLMs for languages such as Basque and Spanish. The research line presented in this project will help to strengthen HiTZ center by placing it at the center of research on LLMs from a multilingual and multidomain perspective in low resource scenarios.

5.4. Plan for scientific communication and internationalization of the results

The PI and host group have a strong track record of publishing at national and international level and they will continue work in disseminating results (both research and application related) throughout the duration of the project. This will include the publication of top-ranking open access journal articles and conference proceedings as well as presentation of the project results at scientific events, workshops and conferences. Furthermore, we will target scientific international evaluation campaigns such as SemEval and CLEF. In addition to the standard scientific means of dissemination, we plan to make use of a website and modern social media such as LinkedIn and Twitter, which offer a cost-effective and quick way of transmitting information.

5.5. Dissemination plan of the results to society

The produced new software will be distributed under open source licenses, enabling the universal access to a new cutting-edge technology in NLP. The feasibility of the socio-economic impact is boosted by the socio-economic and scientific impact that linguistic tools and resources already by the PI the host research group. Examples of this noticeable impact include thousands of downloads of the Multilingual Central Repository (MCR), the linguistic processors such as IXA pipes or our language models uploaded into HuggingFace

repository.[18]

5.6. Summary of the data management plan

DeepMinor will follow the FAIR principles,[19] that is findable, accessible, interoperable and re-usable. At least the following points will be addressed: (i) Dataset description: We will at least collect / generate news, company data, social media data, blogs, images, videos both for extrinsic and intrinsic evaluation. (ii) Data sharing: the ways data will be exploited, shared, and/or made accessible for verification and re-use. We will take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate data. If such access cannot be granted, we will provide the reason for not giving access to specific parts of the research data. (iv) Archiving and preservation: the ways data will be curated and preserved. We will ensure long-term preservation of the data. (v) The data will be documented and described through open access publications and through the project's website and as far as possible through data sharing platforms such as Huggingface or Zenodo.

5.7. Results transfer and valorization plan

The transfer of results plan will consist of the following: Software components will be offered and distributed under open-source licenses such as Apache 2.0, whereas corpora and other linguistic resources will be distributed, whenever possible, under the Creative Commons licensing.

6. Scientific Technical and Training Context

The Doctoral Programme in Language Analysis and Processing falls within the area of language technologies. This area has undergone a major expansion worldwide in recent years, especially since the increase in popularity of applications such as machine translation (translate.google.com), voice communication with smartphones (Apple) and Google's incorporation of the above technology to improve their search results. During this period, speech processing, machine translation and the searching and classification of documents have managed to enter the group of applications of everyday use for regular users of communication technologies. A significant improvement has also been achieved in the interaction between the devices and the people who use them. Our doctoral program aims to prepare researchers to be able to meet these new technological challenges.

HiTZ currently develops and uses the latest language technology including high performance facilities (including the latest 12 A100 GPUs with 80Gb each) and infrastructure for deep learning large pre-trained language models. The PhD students will receive a complementary education through the project itself, complementary courses, internal seminars we organize weekly at IXA[20] and the webinars we organize monthly at HiTZ.[21] DeepMinor researchers also maintain an intense and fruitful collaboration among them and also with Elhuyar and Vicomtech technological centers and other national and international research groups worldwide. HiTZ is also a member of CLAIRE and participates in the TAILOR project. The call for hiring the student will be announced among our Master and PhD students and the research network of all the research groups we collaborate with. HiTZ is also very active in social networks and the Spanish Research networks PLN.NET and the research association SEPLN.

18 https://huggingface.co/HiTZ
19 https://www.go-fair.org/fair-principles/
20 http://www.hitz.eus/en/node/248
21 http://www.hitz.eus/en/webinars

References

Agerri, Rodrigo, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. "Give your Text Representation Models some Love: the Case for Basque." In LREC 2020.

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernández de Landa, Álvaro Rodrigo (2021). VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection. Procesamiento del Lenguaje Natural, 67, pp 173-181.

Rodrigo Agerri, German Rigau (2016). Robust multilingual Named Entity Recognition with shallow semi-supervised features. Artificial Intelligence, 238 (2016) 63-82.

Rodrigo Agerri, Josu Bermudez and German Rigau (2014): IXA pipes: Efficient and Ready to Use Multilingual NLP tools. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), 26-31 May, 2014, Reykjavik, Iceland.

Rodrigo Agerri and Eneko Agirre (2023). Lessons learned from the evaluation of Spanish Language Models. Procesamiento del Lenguaje Natural (70), pp 157-170. https://doi.org/10.26342/2023-70-13.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5799–5811.

Armengol-Estapé, Jordi, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor González-Agirre, Maite Melero, and Marta Villegas. "Are Multilingual Models the Best Choice for Moderately Under-resourced Languages? A Comprehensive Assessment for Catalan." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4933-4946. 2021.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. arXiv preprint arXiv:2111.10952, 2021.

Isabelle Augenstein. 2021. Towards explainable fact-checking. ArXiv, abs/2108.10274.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding.. In EMNLP 2016.

E. Barba, L. Procopio, C. Lacerra, T. Pasini, R. Navigli. Exemplification Modeling: Can You Give Me an Example, Please? IJCAI 2021.

Rishi Bommasani, et al.. On the opportunities and risks of foundation models, 2021. https://arxiv.org/abs/2108.07258.

Tom B. Brown, et al. Language models are few-shot learners. Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020.

Canete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. "Spanish pre-trained bert model and evaluation data." Pml4dc at iclr 2020 (2020): 2020.

Catherine Chen, Kevin Lin, Dan Klein. Constructing Taxonomies from pre-trained Language Models. NAACL 2021.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12:2493–2537, 2011.

Colon-Hernandez, Pedro, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. "Combining pre-trained language models and structured knowledge." arXiv preprint arXiv:2101.12294 (2021).

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised Cross-lingual Representation Learning at Scale." In ACL 2020.

F.-J. Chang, M. Radfar, A. Mouchtaris, B. King, and S. Kunzmann, “End-to-end multi-channel transformer for speech recognition,” 2021.

Yi-Ling Chung, Marco Guerini and Rodrigo Agerri (2021). Multilingual Counter Narrative Type Classification. In Argument Mining 2021.

Chung, Y., Kuzmenko, E., Tekiroglu, S.S., & Guerini, M. (2019). CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In ACL 2020.

Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. Annual Conference on Neural Information Processing Systems (NeurIPS 2015), December 7-12, 2015, Montreal, Quebec, Canada, pages 3079–3087, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL 2019.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning, 2021. Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao

Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, et al. Pre-trained models: Past, present and future. AI Open, 2021

Doddapaneni, Sumanth, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. "A primer on pre-trained multilingual language models." arXiv preprint arXiv:2107.00676 (2021).

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D.Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.

El Baff, R., Wachsmuth, H., Al Khatib, K., Stede, M., Stein, B. (2019). Computational argumentation synthesis as a language modeling task. In INLG, p. 54–64, ACL.

Ecker, U.K., Hogan, J.L., & Lewandowsky, S. (2017). Reminders and Repetition of Misinformation:: Helping or Hindering Its Retraction? Journal of applied research in memory and cognition, 6, 185-192.

Ecker, U.K., O'Reilly, Z., Reid, J.S., & Chang, E.P. (2019). The effectiveness of short-format refutational fact-checks. British Journal of Psychology, 111, 36 - 54.

Eshet,Y (2004) Digital literacy: A conceptual framework for survival skills in the digital era. J. Educ. Multimedia Hypermedia 13, 93–106

Steven Y. Feng, Jessica Huynh, Chaitanya Narisetty, Eduard Hovy, Varun Gangal. SAPPHIRE: Approaches for Enhanced Concept-to-Text Generation. INLG 2021

Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." Minds and Machines 30, no. 4 (2020): 681-694.

Gadetsky, Artyom, Ilya Yakubovskiy, and Dmitry Vetrov. "Conditional Generators of Words Definitions." In ACL 2018.

Sebastian Gehrmann, et al. The GEM benchmark: Natural language generation, its evaluation and metrics. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 96–120.

Golovanov, S., Kurbanov, R., Nikolenko, S., Truskovskyi, K., Tselousov, A., and Wolf, T., Large-scale transfer learning for natural language generation. ACL 2019.

Han, Xu, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu et al. "Pre-trained models: Past, present and future." AI Open (2021).

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. DeepLearning. MITPress, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabel Augenstein. 2021. Few-shot cross-lingual stance detection with sentiment-based pre-training. ArXiv, abs/2109.06050.

Horrigan J.B., Digital readiness gaps (2019). Pew Research Center

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In ACL 2018.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023. Mistral 7b. arXiv preprint arXiv:2310.06825 .

Hua, X., Hu, Z., and Wang, L. (2019). Argument generation with retrieval, planning, and realization. In ACL 2019. .

Huguet, A., Kavanagh, J., Baker, G., Blumenthal, M.S. (2019) Exploring Media Literacy Education as a Tool for Mitigating Truth Decay (RAND Corporation)

Jones-Jang, S. M., Mortensen, T., Liu S. (2019) Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. Am. Behav. Sci.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B.S., Adib, E., Zarka, J., Traboulsi, C., Akl, E.W., & Baddour, K. (2020). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. Cureus, 12.

Dilek Küçük and Fazli Can. 2020. Stance Detection: a Survey. ACM Computing Surveys (CSUR), 53(1):1–37.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In ACL 2020.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. pre-trained language model for text generation: A survey. In IJCAI 2021.

R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," 2020.

Liu, Pengfei, 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." arXiv preprint arXiv:2107.13586 (2021).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742, 2020.

Ma, Kaixin, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. "Knowledge-driven data construction for zero-shot evaluation in commonsense question answering." In 35th AAAI Conference on Artificial Intelligence. 2021.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013a. URL https://arxiv.org/abs/1301.3781.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC2018), Miyazaki, Japan, 2018.

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 preprint arXiv:2111.01243, 2021.

Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E., 2023. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.

Park, C., Yang, W., and Park, J. (2019). Generating sentential arguments from diverse perspectives on controversial topic. In Workshop on Natural Language Processing for Internet Freedom, p. 56–65, ACL.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In EMNLP 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In NAACL 2018.

Peters, Matthew E., Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. "Knowledge Enhanced Contextual Word Representations." In EMNLP 2019.

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, 21:1–67, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever. Improving language understanding by generative pre-training. Technical Report. Open AI., 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

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, 21:1–67, 2020.

Ravichander, Abhilasha, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. "On the systematicity of probing contextualized word representations: The case of hypernymy in BERT." In Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, pp. 88-102. 2020.

Richardson, Kyle, and Ashish Sabharwal. "What does my qa model know? devising controlled probes using expert knowledge." Transactions of the Association for Computational Linguistics 8 (2020): 572-588.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8:842–866, 2020. doi:10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54.

Sainz O. and Rigau G. Ask2Transformers: Zero-Shot Domain labelling with Pre-trained Language Models. Proceedings of the 11th Global WordNet Conference (GWC 2021). Pretoria, South Africa. 2021a.

Sainz O., López de Lacalle O., Labaka G., Barrena A. and Agirre E. Label Verbalization and Entailment for Effective Zero-and Few-Shot Relation Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 1199–1212, 2021b.

Victor Sanh, et al. Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207, 2021.

Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." In ACL 2021

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al., 2023a. Large language models encode clinical knowledge. Nature 620, 172–180.

Stab, C., Miller, T., Schiller, B., Rai, P., & Gurevych, I. (2018). Cross-topic Argument Mining from Heterogeneous Sources. In EMNLP 2018.

Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200, 2020.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T., 2023. Llama 2: Open foundation and fine-tuned chat models. ArXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Annual Conference on Neural Information Processing Systems (NeurIPS 2017).

David Vilares, Marcos García, and Carlos Gómez Rodríguez. "Bertinho: Galician BERT Representations." Procesamiento del lenguaje natural 66 (2021): 13-26.

Xiong, G., Jin, Q., Lu, Z., Zhang, A., 2024. Benchmarking retrieval-augmented generation for medicine. arXiv preprint arXiv:2402.13178

Ke Wang and Xiaojun Wan. 2018. Sentigan:Generating sentimental texts via mixture adversarial networks. In IJCAI, pages 4446–4452

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021. URL https://arxiv.org/abs/2109.01652.

Thomas Wolf, et al. Transformers: State-of-the-art natural language processing. In EMNLP 2020.

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023. Pmc-llama: Towards building open-source language models for medicine. arXiv:2304.14454.

Xue, Linting, 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 NAACL 2021.