

NORA.LLM

Norwegian large language models
as infrastructure



The following document is part of a proposal submitted to the Research Council of Norway under the call *National Financing Initiative for Research Infrastructure*. The proposal's total budget is 212 million Norwegian kroner (NOK 212M), of which NOK 62M is in-kind contributions from the partners. Project partners include National Library of Norway, University of Bergen (UiB), Sigma2 AS, University of Agder (UiA), Norwegian Computing Center, University of Stavanger (UiS), The Directorate of e-health, SINTEF AS, Oslo University Hospital (Cancer Registry of Norway), University of Oslo (UiO), Østfold University College, Simula Research Laboratory AS, NORCE Norwegian Research Centre AS, UiT The Arctic University of Norway, Norwegian Labour and Welfare Administration (NAV) and Western Norway University of Applied Sciences. Norway's two Centers of Excellence for AI (SFFs), SFF Integreat and SFF Digital Narratives, are supporting the project, so are two Norwegian centers for Innovation in AI (SFIs), SFI Visual Intelligence and SFI Media Futures, as well as the large Cofund programmes DSTrain and LEAD AI.

I think the best way to think about this is to think about the type of AI systems that have been released so far as being basic infrastructure in themselves

...

AI is going to become a common platform and because it's a common platform it needs to be open source, if we want it to be a platform on top of which a whole ecosystem can be built.

- Yann LeCun in The US Senate hearing on Artificial Intelligence, 19.09.23

We believe it is of paramount importance that Norway establishes a broad national effort for large language models as infrastructure, with focus on our own language and with complete control of the input data. This ensures the preservation of the Norwegian language, Norwegian cultural values and democratic principles, and it may ensure a safe and secure governing model. Many future research projects will strongly rely on such foundational models, and the Norwegian public sector and industry will also build services on top of large language models. This infrastructure, and its coordinated expertise, will strengthen Norway's position in international research and enhance our capabilities for innovation. The project will also develop competence and ensure that Norway has the trained workforce necessary to confront global AI competition.

There are several good environments for language technology research in Norway. According to the independent *Mainland Scandinavian NLU Benchmark* run by Danish researchers, University of Oslo's NorBERT and National Library of Norway's NB-BERT-large are the two top-performing LLMs for Scandinavian natural language understanding (per November 21, 2023). These groups also have extensive experience in GPT-like models. As part of the Horizon Europe HPLT project, these partners have already compiled data for the largest fully open decoder language models for Norwegian, with training scheduled to start already November 2023. The Language Technology Group at UiO has worked for decades within this research field, building up data sources and educational programs within the field, long before the recent GPT hype. The National Library has been a world-leading environment for digitalization, starting its digitizing programme already in 2006, ensuring access to a massive corpus of Norwegian digitized text.

Computing resources are of importance for building a language model infrastructure. We will establish a pipeline for training of the models and hardware for running the models. The services developed in The Norwegian AI Cloud project will ensure a smooth transfer and integration with the national e-infrastructure provided by Sigma2. Although Norway has

decent hardware and physical infrastructure, we are currently not set for training the largest models. However, through Sigma2 we will utilize Norway's membership in The European High-Performance Computing (EuroHPC) Joint Undertaking, utilizing Europe's most powerful supercomputer, LUMI. The Language Technology Group at UiO has already been selected with a pilot project on LUMI and with data stored on Sigma2's National Infrastructure for Research Data (NIRD).

NORA.LLM is an inclusive project. We are open for including other relevant Norwegian environments for language technology, and we will work closely also with relevant Nordic initiatives. We publish this proposal because we believe in openness and transparency, and we appreciate feedback on how to improve the project plan. The evaluation through the Research Council of Norway will take time, and with the rapid development in the field we will surely adapt the project plan before the project may be able to start.

The NORA.LLM project will use an agile project management approach to ensure that we are adaptable to rapid change in the field. The project will have two phases: a developmental phase (years 1-3) and an operational phase (years 4-10). The main tasks that will be solved in the project are: Legal and copyright issues related to the use of data and open models; compiling Norwegian training datasets of sufficient size and quality; development of foundation models that are aligned with Norwegian values and cultures; enabling systematic benchmarking to assess both capabilities and limitations of the models; ensuring democratic and inclusive processes; optimization of models to solve specific tasks with a specific focus on health and education; developing business models for the post-project exploitation; upgrading existing High-Performance Computing (HPC) infrastructure in Norway. These tasks will be solved across 14 work packages led by Klas Pettersen (NORA.ai), Torger Kielland (UiB), Per Erik Solberg (National Library of Norway), Hans A. Eide (Sigma2), Per E Kummervold (National Library of Norway), Andrey Kutuzov (UiO), Per-Arne Andersen (UiA), Krisztian Balog (UiS), Samia Touileb (UiB), Erik Velldal (UiO), Michael A. Riegler (Simula), Klaus Johannsen (NORCE) and Asbjørn Følstad (SINTEF). The details of the project are described in the full application (144 pages, including the work packages).

The work package leaders have been working hard to establish this collaboration. We would additionally like to acknowledge the contributions of the following individuals to the application: Stefano Nichele, Turgay Celik, Steven Hicks, Andrei Kutuzov, Tore Lunde, Torgar Kielland, Alexander Selvikvåg Lundervold, Asbjørn Følstad, Ole Kristian Våge, Hilde Margrethe Lovett, Benjamin Ricaud, Svein Arne Brygfeldt, Jan F. Nygård, Tom Ryen, Cathrine Pihl Lyngstad, Jon Vegard Sparre, Lilja Øvrelid, Robert Jenssen, Pierre Lison, Morten Goodwin, Yngvil Beyer, Sabry Razick and Gard Thomassen.

A special thanks goes to Michael Solvang, research coordinator at NORA.ai, for excellent coordination of the project proposal!



Klas Pettersen

Contact: Klas Pettersen, klas.pettersen@nora.ai, phone: +47 950 50 294.

Reference to mandatory outline submitted 27 May 2020 no: 347002

1 National importance and relevance to the call

1.1.1 National Importance

Turing award winner Yann LeCun was recently (19.09.23) speaking in the US Senate hearing on Artificial Intelligence, where he discussed AI and large language models (LLMs): “I think the best way to think about this is to think about the type of AI systems that have been released so far as being basic infrastructure in themselves ... AI is going to become a common platform and because it's a common platform it needs to be open source, if we want it to be a platform on top of which a whole ecosystem can be built”.

LLMs have already demonstrated tremendous power, with a vast repertoire of applications enabled by their ability to understand and generate natural language. They have proven to have impressive capabilities for open-ended problem-solving simply through dialog-based interaction. The models can also be optimized for specific tasks through further fine-tuning. This adaptability to a wide range of downstream tasks is why they are known as *base* models or *foundation* models. These models are highly versatile and can be used across various domains, such as healthcare, education, government, public services, and more. Given the transformative potential of these models, they will undoubtedly have a profound impact on society.

However, these models also come with evident challenges. They exhibit biases, harbor cultural preferences inherited from the training data, and may impact political discourse and decision-making. Currently, a handful of companies control the most widely used LLMs, which raises concerns regarding transparency, GDPR, monopoly, and the ad-based business model behind these companies. Interaction with these models is typically only provided via an interface, users cannot access the model itself, details about the training data and the training procedure are not disclosed or the license restrict the use. Further, the current concentration of power regarding LLMs poses challenges to regional digital sovereignty. To address these issues, several nations recognize the need to develop their own national language models. For example, in the spring of 2023 the UK announced a **£100 million** initial investment to establish a [Foundation Model Taskforce \[1\]](#) (later called Frontier AI Taskforce). Similar initiatives are underway in Germany, France, and China, and other countries are expected to follow.

Over the last year, it has become clear that the LLMs should be considered basic infrastructure for society. Many future research projects will be built upon this infrastructure, as well as innovation projects and public sector and industry services. Thus, it is of national importance that Norway is fulfilling what we conceive of as a public responsibility to build LLMs from the ground up, for our own language, with full control of the input data. We need a national infrastructure – comprising both the models themselves and the physical hardware needed to train and deploy them – to ensure that Norwegian culture and democratic principles are preserved, and for Norway to strengthen its international research competitiveness and innovation capabilities. We will develop LLMs that will be provided both as a platform service and as open-access models. The models will be accessible for research purposes and can be deployed by academic institutions, the public sector, Norwegian businesses and startups, and the general public.

1.1.2 Relevance to existing national agendas and strategies

The proposed LLM infrastructure project is strongly aligned with existing national agendas and strategies. The Norwegian prime minister recently announced that the government will **invest 1 billion in AI research**. This implies that we will seek to attract and educate more researchers in the field of AI, and the infrastructure described here is therefore perfectly timed and much needed. Further, the new **Ministry of Digitalization and Governance** was recently established in Norway, and the government is in the process of developing a new national digitalization strategy, where generative AI is assumed to be central. The project also closely aligns with the national strategy for Artificial intelligence [2]. The national strategy has a subsection dedicated to language, *Section 2.2. Language data and language resources*, stressing the importance of good language resources and models in the Norwegian languages.

1.1.3 Relevance to the Long-Term Plan for Research and Higher Education

The project supports Norway's Long-Term Plan for Research and Higher Education, and contributes to all three primary goals of the LTP [3]:

- 1) ‘*Strengthen competitiveness and innovation capability*’: Since 2006, the National Library of Norway has digitized a substantial portion of its books, newspapers, and journals, resulting in **one of the largest**

collections of digitized texts globally. The infrastructure built in this project will take advantage of this investment, and it will be available to the private and the public sector and start-up companies.

- 2) *‘High quality and accessibility in research and higher education’*: Open state-of-the-art Norwegian LLMs will be a resource for researchers, both to enhance their research output and for them to conduct research on. The models will be fine-tuned for education to enhance the quality of education. NORA is already hosting the national research school in AI, where this will be deployed.
- 3) *‘Environmental, social, and economic sustainability’*: The majority of computing will be performed on national Sigma2 e-infrastructure resources, including LUMI, currently the most powerful supercomputer in Europe. Waste heat energy from the LUMI data center accounts for about 20% of heating in the city of Kajaani. Similarly, major computing resources within Norway are located in the Sigma2 data center Lefdal Mine, where waste heat is used for food production. Both data centers are 100% hydro-powered.

The project further aligns with certain thematic prioritizations in the LTP, particularly ‘Enabling and industrial technologies’ and ‘Health’. Artificial Intelligence is highlighted as a key area and Norway has a large amount of high-quality health data, and in this project we will fine tune models into the health domain (WP13). An earlier initiative from NORA exemplifying “AI as basic infrastructure” can be found in box 3.11 in LTP. When the protein folding algorithm **AlphaFold** by DeepMind was open sourced, NORA together with USIT (University Center for Information Technology, UiO) and Sigma2 initiated a swift installation on Norwegian infrastructure. A large workshop was organised, where new proteins important for drug development were shown on the fly, and the algorithm was made available to the research community.

1.1.4 Relevance to Sub-Areas in the Norwegian Roadmap for Research Infrastructure

Sharing and reuse of research data [5]: LLMs are the prime examples of how to reuse research data and data in general. As mentioned, Norway has a very rich repository of digital data, much thanks to years of effort by the National Library, our national health registers, and more. This project is an excellent opportunity to harvest value from these data sources.

Information- and communication technology [5]: The research infrastructure will be at the forefront of digital technological advancements in AI and language processing, which are fundamental to Norway's digitalization efforts. It will ensure digital sovereignty, relying on national computational resources and one of the world's fastest supercomputers (LUMI). It promotes innovation and new AI applications, paving the way for new research and more efficient and intelligent digital services across various sectors.

Life science and health [5]: Natural language processing and large language models have the potential to revolutionize healthcare by enabling more accurate patient diagnosis, treatment recommendations, and the extraction of valuable insights from medical records. LLMs also have the potential to suggest relevant diagnostic codes, etc. We will have dedicated efforts to fine-tuning the models for health (WP13).

Humanities and Social Sciences [5]: LLMs have profound social impacts, influencing communication and information dissemination. **Open access to LLMs is vital** for researchers to understand and address these effects. It allows scholars to scrutinize biases, explore ethical concerns, and analyze the impact of LLMs on societal structures. This access promotes transparency, contributing to responsible development and deployment, ultimately fostering a more informed integration of LLMs into our digital landscape.

Humanities and Social Sciences [5]: LLMs have profound social impacts, influencing communication and information dissemination. **Open access to LLMs is vital** for researchers to understand and address these effects. It allows scholars to scrutinize biases, explore ethical concerns, and analyze the impact of LLMs on societal structures. This access promotes transparency, contributing to responsible development and deployment, ultimately fostering a more informed integration of LLMs into our digital landscape.

2 Excellence

2.1 Contribution to research and innovation

AIM: Develop an AI foundation model infrastructure for which a whole ecosystem of research projects, innovation projects, and services for industry and the public sector can be built.

2.1.1 Objectives

1. Compile large-scale Norwegian datasets for training and testing LLM: This includes balanced text corpora for pre-training, but also structured and annotated data for fine-tuning and testing. The data will be open (as far as copyright allows) and transparent (all sources documented).
2. Legal & copyright: Address legal/copyright issues regarding the datasets and open use of the models.

3. High-performance-computing (HPC): procurement of hardware to accommodate future computing needs for public/business and research requirements.
4. Development of LLMs for Norwegian: Optimize and refine the models through pre-training, chat optimization, and retrieval while maintaining rigorous ethics and alignment standards and benchmarking the performance of the models. The foundation models will be continuously improved during the operational phase by existing and new research projects. The project will draw on, and contribute to, the open-source ecosystem around LLMs.
5. Business Model: Develop a business model that enables open and sustainable use of the models for research purposes, as well as domain-specific fine-tuning for the business sector and public sector.

2.1.2 Research challenges

The current LLM landscape is dominated by models trained mainly on US data. The knowledge and expertise are concentrated in the hands of a few large corporations, which also control the HPC needed to operate these models. This raises issues regarding secure data handling and GDPR. Moreover, in these commercial systems, both the training data and resulting models are essentially “black boxes”, not accessible for inspection or scrutiny. The fact that the software used for training is typically closed-source, further hinders replicability. There are ongoing initiatives in Norway to build generative language models, but these are fragmented efforts carried out by individual institutions in isolation. However, to develop a state-of-the-art infrastructure (data, legal, models, computing, business) of the size we propose is highly ambitious; to achieve this, we believe that a large-scale national collaboration is required.

2.1.3 Contribution to excellence in science, knowledge needs and innovation

The development of the LLM infrastructure will be a national joint effort and will significantly enhance knowledge and expertise within language technology and related fields, which is pivotal for Norway's digital transformation and safeguarding digital sovereignty. Open LLMs, trained on extensive Norwegian datasets, will facilitate researchers to understand and apply these models within the unique cultural context and the ethical and legal principles of Norway. The foundation models can be fine-tuned towards various domains and sectors (e.g., healthcare and education), enabling innovation and research and thereby enhancing productivity and data-driven insights. By further developing High-Performance Computing capacity within Norway, we can ensure adequate computational resources and secure data management - an inherent challenge with LLMs.

2.2 Quality and uniqueness of the research infrastructure

2.2.1 Landscape of the Infrastructure

The proposed LLM infrastructure will operate within a complex, interconnected, and fast-changing landscape. With their vast repertoire of capabilities, LLMs will influence most industries and sectors, including Healthcare, Education, Legal, Science & Technology, Media & Entertainment, and Government & Public Services. Notable Nordic LLM initiatives include NorGPT by NTNU/NorwAI (Norwegian Research Center for AI Innovation) [6] and GPT-SW3 by AI Sweden [7]. NTNU's centre working on NorwGPT, NorwAI, has several areas of focus, one of which is language technology. However, we strongly believe a national approach to such an ambitious project is needed, with more solid funding, data collection strategy and computational power. AI Sweden's GPT-SW3 is more advanced than NorGPT, however, linguistically the focus is Swedish. Internationally, the landscape is even more dynamic, with leading countries and organizations heavily investing in AI research and development. Projects like OpenAI's ChatGPT models, Google's PaLM-2 and BERT, and Meta's Llama models, have set global benchmarks in the field of LLMs. A Norwegian national research infrastructure for LLMs and language technology will be crucial for participation in expected upcoming calls from European framework programs.

2.2.2 Justification of the initiation of the project

These models are foundational in various applications and should be considered Digital Public Infrastructures, much like the Internet, as argued by several world-leading experts, including Yann LeCun in the US senate. They will likely produce economic value that is far greater than their cost, and globally, Norway is well-positioned to excel in the development of LLMs. Our vast digital data corpus, curated by the National Library over the years, is a remarkable asset. Norway also has extensive domain-specific repositories of high-quality data, e.g., within health. Further enhancing our capabilities, Sigma2 and LUMI provide access to sustainable hydropowered high-performance computing (HPC) [4]. Through a collaborative national effort, this project presents a compelling opportunity to harness Norway's rich data

resources and sustainable computing power, and will significantly enhance Norway's knowledge and expertise within the important field of language technology and related subject matters.

2.2.3 Description of the research infrastructure

This will be an unusual and unique infrastructure in that it will be continuously evolving, see Figure 1. These models are advancing at an unprecedented pace, and it is important to ensure that the infrastructure continues to be cutting-edge, also on the operational phase of the project and after the project has ended. In years one to three, state-of-the-art models of various sizes and complexities, that serve diverse needs, will be developed.

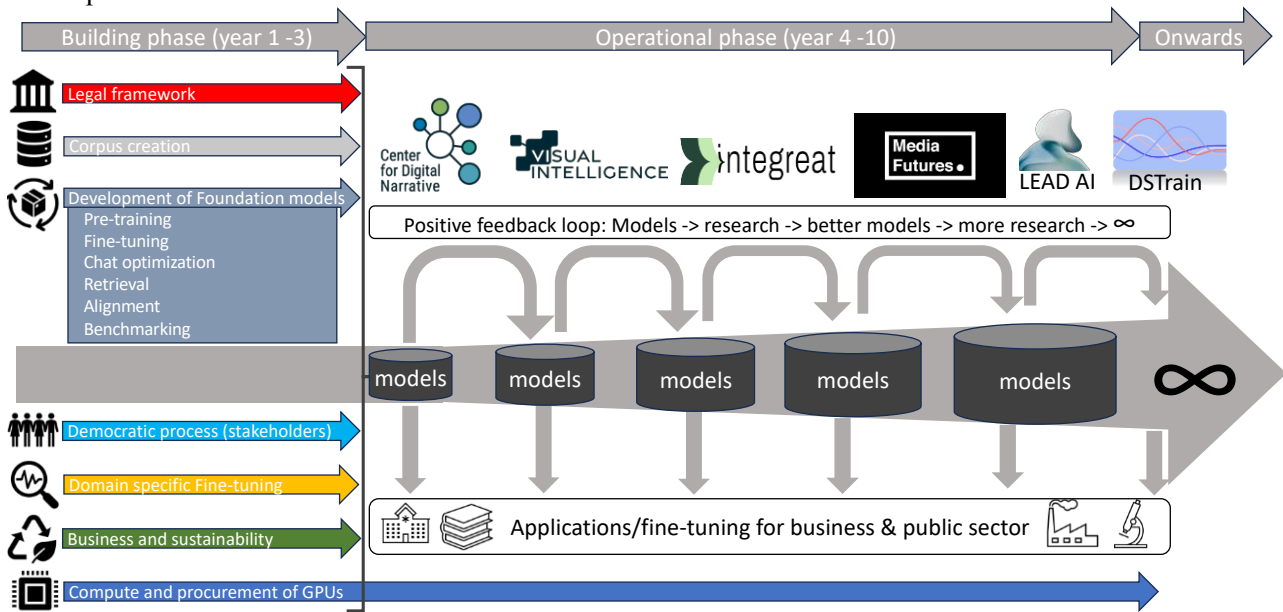


Figure 1 – Overview of the infrastructure, with a positive feedback loop between research and infrastructure.

Due to the rapid advancement in the field and an estimated project start medio 2025, the exact architectures, and whether the models will be built completely from scratch or by further tuning of existing pre-trained models will be decided in the projects initial phase. Data collection (WP4) will be an important aspect of the project. The models will be trained on the most comprehensive corpus of Norwegian data available, reflecting Norwegian culture and values. The models will be open, and the process of development will be transparent and well-documented. Benchmarks to assess natural language understanding capabilities of the models will be developed (WP11) as well as potential ethical challenges will provide crucial knowledge to downstream users. To ensure that the models stay cutting-edge, we have built in a ‘positive-feedback-loop’ for the operational phase and onward. The models will be continuously improved via the effort of existing and new research projects, such as the Centres of Excellence *Digital Narratives* [8] and *Integreat* [9], the Centre for Research-based Innovation *Visual Intelligence* [10], *Media Futures* [11] and the EU cofunded projects *DSTrain* [12] and *LEAD AI* [13] (see letters of intent from these centres/projects). Access to the models will stimulate research, which improves the models, which in turn stimulates more research, in a positive feedback loop. There will be a significant update of high-performance computing in connection with Sigma2 and the Norwegian AI cloud, to ensure sufficient computing power to operate the models.

The outcome will be state-of-the-art Norwegian LLMs, that are trained to encapsulate Norwegian culture, ethical and democratic principles. The models will be offered as platform service (similarly to ChatGPT), and as open models that can be downloaded and fine-tuned to solve specific tasks for businesses and the public sector. Assistance with fine-tuning for businesses and public sector will be offered thorough the European Digital Innovation Hub NEMONOOR.

2.2.4 State-of-the-art models

An open LLM infrastructure, that can safely handle sensitive data, comply with GDPR, and can be used for research purposes, as well as by the business sector and public sector, requires a state-of-the-art ecosystem. **Data Corpus for LLM training** Through our partners, we have access to a very large digital collection of natural language data, as well as high-quality health data and public-service data. These will be compiled into large-scale Norwegian datasets for training and testing LLMs (WP3). This includes balanced text corpora for pre-training, but also structured and annotated data for fine-tuning, testing, benchmarking (WP11), alignment (WP10) and chat-optimization (WP8). The data will be open (as far as copyright allows) and transparent (well documented). NB holds one of the largest corpora in the world stemming from

digitized books, newspapers and journals. In addition, the proposed WebData infrastructure (outline no L347132) will provide fresh and historical web data to NORA.LLM.

Languages and culture The project will establish the most comprehensive corpus of training data in Norway and will ensure that Norwegian language, culture and democratic principles are safeguarded. The model will be trained on the two written languages Norwegian bokmål and Norwegian nynorsk, and we will experiment with the smaller Sámi languages. The models will also be trained on the other Scandinavian languages (Swedish and Danish), as well as English.

Model sizes We will build models of all sizes, small (10M-100M parameters), medium (100M-500M parameters), large (1B-10B parameters) and very large models (10B-100B parameters). Models of different sizes have pros and cons, the smaller ones are faster to train, require less data, are more energy efficient, are suitable for applications where real-time processing is critical and can be used in smaller mobile devices, whereas larger models are needed for complex tasks such as open-ended problem solving and advanced conversational chatbots.

Transparency Foundation models are becoming ubiquitous. Their societal impact is quickly rising whereas transparency unfortunately is not. According to the 'Foundation model transparency index', which measures 10 dimensions of transparency, the most transparent model only scores an average of 57%. [14] The models and pipelines developed in this project will be transparent, documented, and made publicly available.

Computing power and secure data handling To train and operate LLMs, access to HPC – High-Performance-Computing – is essential. We will further develop and use sustainable HPC provided by Sigma2 and LUMI (currently the fastest supercomputer in Europe). This ensures sufficient computing power to both train the models and operate the infrastructure, and safe data handling.

2.2.5 Requirements for e-infrastructure and integration with existing e-infrastructures

This project will establish an e-infrastructure for LLMs, to be included in the greater national e-infrastructure for research and higher education managed by Sigma2. As such, this project will be based on existing components of the Sigma2 and Norwegian AI Cloud national e-infrastructures, and in sum extend the national capabilities without constructing unnecessary parallel structures. The advantages of this arrangement are many, but mainly that this project can build upon what already exists and focus investments on only the missing components for a national LLM infrastructure. Another benefit is taking advantage of the economy of scale and well-established structures for sustainable operation. What is crucial in terms of e-infrastructure requirements is computing resources for model training and inference. The plan for meeting these requirements has two main strategies:

- a) utilize existing and coming EuroHPC capacity suitable for LLM training, such as the LUMI supercomputer, where Sigma2 is a partner,
- b) invest in own national capacity, within Norway and integrated with Sigma2 infrastructure, such as NIRD, for training and inference.

Investments will be based on a thorough analysis of user needs and strategic considerations of costs, technology roadmaps, and market situation (WP5).

The Research Council of Norway has invested substantially in data storage and management implemented by the Sigma2 National Infrastructure for Research Data (NIRD). This project will utilize the NIRD infrastructure for larger storage needs. The data collected or generated by this project will be processed, stored and managed in accordance with the FAIR principles, while ensuring that privacy and security concerns are met. Sigma2 data archiving and management tools, along with tools, policies and strategies developed under the Norwegian AI Cloud, will be used as the starting point for this. Data that includes personal identifiable information will be stored and shared for further analysis using the Services for Sensitive Data (TSD) facility at the University of Oslo where there are Sigma2 computing and storage resources suitable for this purpose.

The national e-infrastructure managed by Sigma2 is primarily located in a state-of-the-art green datacentre, powered only by renewable energy sources and using waste heat for food production. The power-demanding components of the LLM e-infrastructure will be located here or at the above-mentioned international facilities.

2.2.6 Ethical issues and measures to avoid potentially undesirable effects

Engaging, developing, and advancing LLMs, necessitates careful reflection and ethical considerations. One important aspect is responsibility of outputs generated by LLMs, and who should be credited or who have the rights on the generated text. Another main ethical concern is the ability of LLMs to perpetuate stereotypes (as gender bias) and unfair discrimination; and generate toxic language, express hate and use of derogatory language, promote (self-)harm, and generate content inciting sexual or violent acts. In order to facilitate progress in responsible innovation, it is imperative to get a profound comprehension of the potential

harms presented by these models. To this end, we will establish guidelines and norms to ensure transparency and responsibility, we will also provide and develop approaches to mitigate the perpetuation and amplification of harmful content generation. These considerations will be focused on in WP3 and WP10.

LLMs have the potential to generate both positive and negative impacts on society. The proposed project has a clear impact on climate, the environment, and society at large. Training LLMs can result in significant carbon emissions, which can have disastrous impacts on a larger scale. In this project, and in collaboration with Sigma2, all training will be made on infrastructures assuring that waste heat will either be used to heat cities or fuel food production, reducing the negative impact on climate and reducing the carbon footprint of the trained models. Further, WP12 will ensure a good democratic process when building the models, by creating an inclusive AI review and policy-making platform.

3 Impact

3.1 Potential impact of the proposed research infrastructure

3.1.1 Excellence in research

There are still many open questions and challenges regarding LLMs and multimodal models. How do they learn from large amounts of unstructured and heterogeneous data? How do they generalize to new tasks and domains? How do they perform reasoning? How can we evaluate their capabilities and limitations? How can we ensure their alignment with human values and expectations? How do they affect society and the environment? What are the ethical, legal, and social implications of their use and misuse? Given the impact LLMs already have on society, these questions are among the most important questions of our time.

This infrastructure will be invaluable for researchers to study and answer these questions in a Norwegian context. We will develop a portfolio of language models specifically tailored for Norwegian text in various sizes and complexity, which will be made available for research purposes. Transparency will be at the core of the infrastructure, and researchers will have access to the training data and detailed model cards, including benchmarking results against a suite of new Norwegian LLM benchmarks. We will work closely with existing and new research projects, such as *SFI MediaFutures*, *SFF Integreat*, *SFF Digital Narratives*, *SFI Visual Intelligence*, and the COFUNDs *DSTrain* and *LEAD AI*, and provide them with access to cutting-edge technology.

Taken together, this project will have large positive impacts on science and excellence in research. It will enhance the expertise in most disciplines relevant to LLMs and multimodal models, which will be indispensable for Norway. Equally important will be the national effort and resulting eco-system; the knowledge of how to run and operate these models with all their implications.

3.1.2 Recruitment to science

LLMs and foundation models have gained massive attention after the release of ChatGPT, and generative AI is the hottest and among the most important research fields right now. This attention alone makes the infrastructure very desirable, and if Norway can provide a proper infrastructure (data, range of models, required computing power, user interface, legal framework) this will attract students, researchers, and businesses. We expect such an infrastructure together with the AI focus from the government (NOK 1 billion going into AI research) may also attract world leading researchers to Norway. We know that in the coming years, Norwegian universities will have a large influx of PhDs/postdoctoral researchers (100+ students via the projects *Integreat*, *Visual Intelligence*, *DSTrain*, *LEAD AI* alone) within the field of artificial intelligence. Many of these young researchers can take advantage of the infrastructure we are developing.

3.1.3 Internationalization of Norwegian science

For Norway to remain internationally competitive, and to ensure that our culture and democratic principles are safeguarded in the digital world, this infrastructure is timely and required. A combined national effort will contribute to increased recruitment to the field of NLP in Norway and the proposed infrastructure will attract international students, researchers, and businesses. We will build on our Nordic expert group of people working on similar projects in the Nordic countries, to strengthen the Nordic consortium on language models. Furthermore, NORA already has a MoU with The Alan Turing Institute (ATI), Norwegian PhDs are already enrolled into the ATI Enrichments Scheme as the only country outside UK, and several joint research projects have been established. With our infrastructure for LLMs we plan to strengthen this link to the Nordic countries and UK.

Internationally, one of Norway's main competitive advantages is likely within healthcare, due to our extensive repositories of high-quality health data. The proposed infrastructure will enable further use of this

data for research and innovation. On the European level, initiatives are to be expected from programs such as Horizon Europe and the Digital Europe Program, where the chances for Norwegian participation, e.g. in consortia for calls on LLM/NLP, will be greatly increased based on this infrastructure.

3.1.4 Contribution to societal challenges and UN Sustainable Development Goals

Large language models have the potential to positively contribute to societal challenges and UN Sustainable Development Goals (SDGs). This infrastructure will enable research and development that likely will be relevant for several UN SDGs, but the two main focuses in this project are health and education, UN SDG 3 and 4, respectively. The foundation models will be fine-tuned on data from various sources, such as doctors' notes, patient reports, blood results, health registers, etc. to develop more efficient and accurate healthcare systems. Pilots within this field are already started at some of the partner institutions. Two of the project partners, The directorate of eHealth and The Cancer Registry of Norway will ensure secure access to the required data for fine-tuning towards health.

However, it is important to note that large language models also have limitations and potential negative impacts on society. For example, they may perpetuate biases or reinforce existing power structures if not designed and implemented carefully. As mentioned above, it is crucial to consider the ethical implications of these technologies and ensure that they are developed and used responsibly. To ensure the safe and legal deployment of these models we have WPs specifically dedicated to legal issues and ethical alignment (WP3, WP8, WP10 and WP12).

3.1.5 Value creation and national competitiveness

The smaller models will lower the entry barrier for external parties interested in leveraging the technology and stimulate the use of “dormant data” to enhance innovation and productivity across sectors. To ensure the safe and responsible use of data, we will establish a legal framework for large language models and sharing of training data in a Norwegian context (WP3). Services for secure handling of sensitive data and assistance with fine-tuning will be offered through existing services for sensitive data in Sigma2 and USIT (UiO).

3.2 Measures for communication and utilisation

NORA is an inclusive Norwegian research consortium consisting of 17 university and institute partners, and a natural gateway to the Norwegian AI community. NORA was established as a consortium to strengthen collaboration and communication within the field of AI. NORA is also the host of the national research school in AI and the innovation network NORA.startup. NORA's information channels will be actively used in the dissemination of the infrastructure and its related services. NORA has a broad outreach to universities, research institutes, students, startup companies, and incubators. The infrastructure will be central to many of Norway's centres of excellence (the SFIs and SFFs mentioned above), and these centres' communication channels will also be actively used in the dissemination of the infrastructure. For public dissemination NORA's newsletter already reaches about 2500 individuals, and NORA's social media channels (LinkedIn, Twitter/X, Facebook) has a broad impact, strengthened by all our public appearances and conferences, such as NORA's quarterly CuttingEdgeAI conference for industry, public sector and the general public.

The infrastructure will be released as a platform service and as open models that can be downloaded. For the platform, the user will have access via a user interface or interfaces where the prompting and inference take place. Given the pace at which these models are developing, it is difficult to predict the exact computing needs, but we estimate 1.000 to 10.000 active users/month when the infrastructure is in the operational phase (see justification for this assumption in chapter 4.1.5).

We are in the early days of LLM deployment, and ChatGPT will change the business model of many of the larger companies. Therefore, we have dedicated WP14 to develop a sustainable business model for the infrastructure. We will develop a variety of models of different sizes and complexity, which can be used for research purposes or fine-tuned to solve specific tasks in the public sector or by businesses. These models will be downloadable from the NORA website with the accompanying model card and from the repository of the Norwegian Language Bank, and services for deployment of the models will be offered through the European Digital Innovation Hub NEMONOOR.

The NORA.LLM website will contain user documentation, access information, status, training event calendar and contact information, and will serve as a portal for the infrastructure. Personnel for user contact and coordination will be interconnected with the existing pool of IT staff that operates the Sigma2 systems. Open text datasets developed in the project will be distributed through a data platform hosted by the Language Bank at the National Library of Norway. In addition to textual data and metadata, the data platform will also provide text sanitization pipelines which can be dynamically applied to the texts. Both the training data and the data generated by the models during operation will be aligned with the FAIR principles. The training data corpus will feature machine-readable metadata, globally unique and persistent identifiers,

and extensive documentation on data provenance to enhance findability (Findable). Open access will be granted for data without privacy or copyright restrictions, while entry to restricted datasets will require authentication (Accessible). Distribution will occur through an API, utilizing standard exchange formats like JSONL for the data, and metadata will be constructed using standard vocabularies to facilitate interoperability (Interoperable). To ensure the corpus is well-suited for reuse and replication, both data and metadata will be thoroughly documented (Reusable). Data generated by the models will be accompanied by a reference to FAIR principles.

Generated knowledge on legal, corpus creation, model development, democratic processes, domain specific fine-tuning, business & sustainability, procuring, installing, maintaining, and operating the infrastructure will be documented and shared through reports via our communication channels.

To ensure stakeholder involvement WP12 is dedicated to establishing a democratic and inclusive platform, operationalised as a Responsible Innovation Lab, to deliberate and formulate policy on AI-generated content's limits and ethical use. Grounded in the Living Labs methodology, this initiative will foster real-time co-creation of knowledge among a diverse spectrum of key stakeholders, including technologists, policymakers, civil society representatives, and the general public.

4 Implementation

4.1 Project partners, management and organisation

4.1.1 Consortium description and Project partners

Building and operating this infrastructure is highly ambitious. To achieve the objectives, we believe a national effort is required, with a consortium that has expertise in all fields relevant to LLMs. In this proposal we have established such a consortium. The foremost experts in Norway within language technology, machine learning, corpus creation, legal aspects & copyright, business development & sustainability, and high-performance computing, a total of 17 partners, will work together to realize this infrastructure; seven universities/university colleges (UiA, UiB, UiO, UiS, UiT, HiO, HVL), four research institutions (NORCE, NR, Simula, SINTEF), six institutions from the public sector (Directorate of eHealth, National library, NAV, The Cancer Registry of Norway, Sigma2), with the secretariat of NORA - Norwegian AI research consortium, as a coordinator. NTNU was also invited into the project, they declined the invitation, but both they and we are open to collaborations at a later stage, and we see it as likely that we join forces if this proposal is funded. UiO will be the host institution as this is NORA's host institution. UiO has a proven track record as a host institution, and has the required expertise in language technology and HPC. The consortium will be settled through a mutual agreement between the partners at an institutional board level; this will be done during negotiations with RCN when the proposal is funded.

The project partners are as follows:

NORA - Norwegian AI Research Consortium - is a Norwegian collaboration between 8 universities, 3 university colleges, and 5 research institutes. NORA aims to strengthen Norwegian AI research, education, and innovation. NORA will be the project coordinator, for which the NORA secretariat has extensive experience. The NORA.LLM project is perfectly aligned with NORA's strategy.

UIO-IFI: Language Technology Group (LTG) of the Department of Informatics (IFI) at UiO:

LTG has a long tradition of working with HPC-driven ML-based approaches to NLP, including training and benchmarking of LLMs and curation of datasets. Currently ongoing research collaborations include the Horizon Europe project *High Performance Language Technologies*, the *Integreat* Centre of Excellence, and the Centers for Research-based Innovation *MediaFutures* and *NorwAI*. The proposed project is a natural extension of work currently carried out at LTG and fits very well with the research strategy of the group and the department, in addition to UiO's Roadmap for Research Infrastructure. In the project, LTG will primarily focus on LLM instruction tuning and benchmarking, but also other areas like alignment. In the operational phase, LTG will ensure continued maintenance of the data repositories for fine-tuning and benchmarking.

UiO-USIT provides all central IT-solutions for UiO. For example, USIT developed and has hosted TSD - a national collaborative IT platform for sensitive data research - since 2014. This system now supports 5000+ users within 1000+ research projects from more than 42 institutions. USIT have been / are also members in PRACE, EUDAT, EOSC, EOSC-Nordic, EOSC HUB, EOSC Life, EuroHPC, and several NeIC projects. Further, USIT is supporting several large research centres (CMBN (SFF), CCB (SFF), NCGC, NORMENT (SFF/KG Jebsen), LCBC (EU project), CEES (SFF), CanCell (SFF), and many more). USIT will contribute to WP6 and WP14. NORA.LLM fits well with the UiO Roadmap for Research Infrastructure [15].

The Cancer Registry of Norway (CRN) is an independent research institution under Oslo University Hospital Trust. CRN has systematically curated textual information on cancer patients since 1952, ensuring

completeness and high data quality. This data curation is currently done manually by 20 medical coders and would be greatly improved by utilizing LLM technology. The CRN will contribute by providing 10 years of medical text and be involved with data management for sensitive data and fine-tuning for the health domain.

Simula Metropolitan Center for Digital Engineering (SimulaMet) is a government-owned research institution. SimulaMet has internationally recognized competence on advanced machine learning systems, including work on foundational models such as GPT. Training foundation models is well aligned with Simula's strategic research areas. Simula already hosts the national research infrastructure eX3, which is not well suited for training LLMs. In the project, SimulaMet will lead the work package on the democratic process and will also actively contribute to the work package on domain-specific fine-tuning

SINTEF has research-based expertise on artificial intelligence and machine learning, trustworthy human-AI interaction, sustainable digital systems, and digital innovation. Involved researchers are located at SINTEF Digital. The NORA concept for a Norwegian large language model infrastructure fit well with SINTEF's strategy as a leading provider of digital research and innovation, and SINTEF will facilitate its uptake through our substantial collaboration with industry and government in research and innovation. SINTEF will contribute to benchmarking, domain-specific fine tuning of the infrastructure, human-centred design competency, and will lead the activity on business and sustainability. **University of Agder's** Center for Artificial Intelligence Research (CAIR) houses around 40 full-time researchers, encompassing students to professors, dedicated to both fundamental and applied AI research, with a focus on explainable AI, deep learning, NLP, and RL. This project enhances CAIR's research and engages students in data analytics, adhering to the budget. It aims to improve LLM capabilities via an open, transparent, and public approach, striving to develop infrastructure for optimizing chat functionality in current and future large language models, with a priority on user privacy and inclusive model training.

Western Norway University of Applied Sciences (HVL). The AI Engineering group at HVL specializes in applied artificial intelligence, focusing on healthcare, education, and experimental physics. At the Mohn Centre for Innovation and Regional Development, we have practical experience in applying the concept of "living labs" to foster a responsible co-creation arena of radical technologies with companies, authorities, universities, and the broader society. In this project, we will assist in developing and evaluating AI-based software and also contribute to the WPs on alignment and democratic processes. Post-project, we will use and contribute to the infrastructure developed in the project.

NAV will collaborate with NORA and partners on the proposed research infrastructure on large language models. We intend to participate in five work packages (WP4, WP7, WP8, WP9, WP12) in the project with a total of 22 person months. We will offer the needed competence from us to fulfil our obligations in each work package. Participating in this project is in accordance with NAV's research and development plan (FoU-plan 2021-2025). It has «Interaction with our Users» as a strategic research area, this includes research on both methods and technology for improving how we interact with our users

National Library of Norway (NB) is responsible for WP4 and WP6. NB hosts the Norwegian Language Bank, a national infrastructure for language technology that provides and develops open access datasets for Norwegian. The AI lab at NB is a pivotal contributor to Norwegian NLP and has developed open ML models for both speech and text, with cutting-edge performance. Leveraging the institution's expertise in copyright law, GDPR, mass digitization and data center operations, NB will ensure that both data and models are shared in accordance with the FAIR principles and with the most open licenses possible.

NORCE (Norwegian Research Centre AS) is a leading independent research institute in Norway, conducting research in energy, healthcare, climate, environment, society, and technology. With its expertise in machine learning, AI, computational sciences, and modeling, NORCE will play a crucial role in NORA.LLM. It is part of the NORA network, EuroCC2, and the NAIC project. In NORA.LLM, NORCE will be responsible for fine-tuning the language models for education, health, and industry. Post-project, NORCE will continue to maintain and develop the infrastructure. The project is fully aligned with NORCE's long-term strategy.

The Norwegian Computing Center (Norsk Regnesentral, NR for short) is one of Norway's leading research institutions within machine learning, statistical modelling and ICT. NR has previous experience in corpus creation, fine-tuning of various large language models and adaptation to dialogue use cases, notably by leading the two NFR-funded projects CLEANUP. The three NR researchers involved in the project will be Ildikó Pilán, Mark Anderson and Pierre Lison, all three permanent research scientists specialized in NLP and machine learning. The proposed project is a very good fit for NRs focus on applied R&D within the fields of statistical modelling, machine learning and language technology.

Sigma2 AS has responsibility for and manages the Norwegian national e-infrastructure for large-scale data- and computational science and participation in related European activities such as EuroHPC. Sigma2 owns and operates national HPC systems, including the Norwegian national share in LUMI, the AI-focused pre-exascale EuroHPC system in Kajaani, Finland. Central to Sigma2 services is the storage infrastructure

(NIRD - Norwegian Infrastructure for Research Data) combined with services for the whole research data lifecycle.

University of Bergen (UiB). The Department of Information Science and Media Studies (InfoMedia) and the *MediaFutures SFI* at the Faculty of Social Sciences, the Faculty of Law, and SLATE at the Faculty of Psychology lead a variety of national and international AI-related projects. Each participant has expertise in both developing solutions and providing theoretical insights on LLMs. InfoMedia will lead works on ethics and alignment and the Faculty of Law will lead work on legal framework and data sharing. In addition, participants from UiB will be responsible for providing insights and solutions to corpus creation, benchmarking, retrieval, and fine-tuning of models. The project matches the long-term strategy of UiB.

The Arctic University of Norway: The Machine Learning Group at UiT is a nationally and internationally recognized research group in machine learning and AI. The group heads Visual Intelligence and is a major component in Integreat, which are two Norwegian excellence centers funded by the Research Council of Norway. The group has a strong expertise in deep learning and on all the aspects concerning training and fine-tuning deep learning models. This applies in particular to LLMs and text analysis. Within the project, the Machine Learning group will take part in fine-tuning and alignment of LLMs and will be responsible for a task in fine-tuning for applications (e.g., medical, education, industrial). This project will lay the foundation for an AI cloud-based infrastructure that is well aligned to the group and university strategy.

Norwegian Directorate of e-health (E-health) will contribute to the project with specific competence on linguistic and language technology as well as general and health specific AI competence and experiences. We will participate in WP 4 Corpus creation and WP 13 Domain-specific fine-tuning with a focus on health, as well as WP3 Legal framework, WP 10 Ethics and Alignment, WP 12 Democratic processes, WP 14 Business and sustainability. AI is a central component in the National e-health strategy for the Norwegian health sector and well as one of the main priorities for the Directorate for 2024.

Østfold University College (HiØ) – Artificial Intelligence for a more digitalized society.

Our HiØ team contributes with strong expertise at the intersection of digitalization and society, relying on a strong knowledge on the technological aspects of AI technology, as well as their ethical and societal implications. Our focus will be on the pre-training aspects of the models and the domain-specific fine tuning, where we have extensive expertise withing the education and healthcare domains. Additionally, we will contribute with our strong expertise on alignment and democratic processes, in order to assess and evaluate the impacts of the utilization of LLMs in our society. The scope of this infrastructure project is aligned with HiØ's strategy, particularly with the priority area "The Digital Society".

University of Stavanger (UiS): The Information Access & Artificial Intelligence (IAI) research group within the Dept. of Electrical Engineering and Computer Science has established itself at the forefront of international research on information retrieval and natural language processing. IAI leads the Language and Personalization WP in the NorwAI SFI and also participates in the MediaFutures SFI. As part of the Stavanger AI lab, it is a strategic priority for UiS to bring the research to partners in industry and in the public sector. UiS will lead WP9 Retrieval and contribute to WP4 Corpus creation, WP7 Foundation model fine-tuning, and WP1 Benchmarking.

4.1.2 NORA.LLM governance model and project management:

Advisory board The advisory board will have 5 members, and will officially be appointed by the Managing board. Magnus Salghren (Head of Research AI Sweeden), Serge Belongie (Director Pioneer Center for AI in Danmark), and Peter Sarlin (CEO Silo AI) has already been asked and confirmed their interest in attending. The advisory board will advise the managing Board on main strategic decisions. Adjustments to the board can be made if deemed necessary based on user needs, expertise, etc.

Management board The managing board will consist of one member from each partner institution, plus the project manager. The board will have quarterly meetings, or more frequent if required. The Project manager will be responsible for the operation until handover to Sigma2.

Project manager The project will be coordinated and managed by Klas H. Pettersen (PhD), CEO of NORA. The project Manager will oversee and ensure that the planned WPs are carried out and that the milestones are met. The project manager is responsible for reporting to the research council.

Project leader The project manager and the management board will employ a well-suited operational project leader, who will be responsible for daily operations.

Project working group The group will consist of the WP leaders and/or at least one member from each partner institution and the project leader appointed by the management board.

Reporting lines The reporting lines go from the project working group à project leader à project manager à management board, and back. The advisory board operates on the side of the reporting lines and gives their advice to the managing board. This organisation will avoid unnecessary coordination between the consortium partners and defines the responsibilities of each of the partners clearly.

Procurement Sigma2 will be responsible for the procurement process and overall maintenance and operational program of the computing infrastructure during the project period. The National Library will be responsible for the procurement of hardware to be placed at their location in Mo i Rana.

4.1.3 Expertise and experience of project manager and WP leaders

WP1 and WP2: Klas Pettersen holds a PhD in computational neuroscience (2007). He has held different research positions within neuroscience since he started his PhD in 2001. As a theoretical neuroscientist he has extensive experience in neural networks. Since 2019 Dr. Pettersen has been the CEO of NORA, where his main responsibility has been to coordinate Norwegian research, education and innovation within AI. Dr. Pettersen has also been part of several Norwegian AI committees put down by the Ministry of Education and Research (KD)/The Research Council of Norway, e.g., the (1) Data Infrastructure Committee, (2) the Advisory Expert Group established for Norwegian AI Initiative (“the AI billion”) and in (3) the committee working for KD on establishing an MoU between US Department of energy and KD on collaboration on AI.

WP3: Torger Kielland (PhD) primarily works with intellectual property law and contract law. His research especially focuses on the protection and exploitation of information technology, including software patenting and open software licensing. This background makes him particularly suited to handle issues concerning the legal framework for training and utilizing a Norwegian language model.

WP4: Per Erik Solberg has a PhD in linguistics and has worked with research and development of speech and language technology at the National Library since 2019. He is work package leader in the NFR-funded SCRIBE project on speech recognition of dialogues and dialectal speech. In 2022 he was a member of the expert panel for speech technology at the Norwegian Board of Technology.

WP5: Hans A. Eide holds a PhD in Atmospheric Science with a strong focus on high performance computing. Since he has 2006 been working exclusively on HPC and related technologies, first for the University of Oslo IT department, and since 2016, for the national e-infrastructure provider Sigma2 AS. He currently has a position as Special Adviser and heads Sigma2’s services for Advanced User Support and several international projects and collaborations where Norway is represented by Sigma2. The latter includes the EuroHPC LUMI consortium. Additionally, Hans is on the board of the Nordic e-Infrastructure Collaboration (NeIC) and appointed permanent expert on e-infrastructures to the ESFRI Forum.

WP6: Per E Kummervold (PhD) has developed the NorthT5-models and has in collaboration with EPFL, developed the COVID-Twitter-BERT-model. At the National Library he has been project leader in developing the first working Norwegian transformer model (NB-BERT-Base and NB-BERT-Large) in addition to leading the work with developing and releasing the Norwegian Colossal Corpus. Lately he has lead the development of NB-Whisper.

WP7: Andrey Kutuzov has a PhD in computational linguistics and worked extensively with training and fine-tuning large language models both for Norwegian and other languages (NorBERT and NorBench projects). He serves as Oslo onsite manager of the EU Horizon project “High-Performance Language Technologies”, which aims at training open baseline language models for all European languages.

WP8: Per-Arne Andersen, an Associate Professor in Artificial Intelligence and Cybersecurity at the University of Agder, holds a PhD specializing in reinforcement learning methods. His industrial collaborations have advanced deep learning applications, especially in safety-critical domains. Currently leading NRC-funded projects like SecureIoT, he also acts as a Senior AI Advisor for Sørlandet Sykehus. With a commendable h-index of 8 and 236 citations since 2018, Dr. Andersen has a rich publication record. He’s an active public speaker, adept at disseminating research findings to the public

WP9: Krisztian Balog (PhD) is an internationally recognized expert in the field of information retrieval (IR), with an h-index of 46. Balog serves on the senior program committee of leading international IR and AI conferences, has served as general co-chair of ICTIR’20 and ECIR’22 (both organized in Stavanger, Norway), and as coordinator of various IR benchmarking efforts. He is the recipient of the prestigious Karen Spärck Jones Award for his work in the interconnection of IR and NLP. Balog leads the Language and Personalization work package within the NorwAI SFI.

WP10: Samia Touileb (PhD) has a pivotal role within the context of the NORA-LLM project. Holding a PhD in Natural Language Processing, she has dedicated the past few years to developing and releasing innovative resources and methodologies aimed at addressing and ameliorating the challenges associated with objectionable content generated by Norwegian LLMs. Her distinguished expertise and research focus are essential for steering the work package toward its objectives, ensuring that ethical considerations and alignment of LLMs remain at the forefront of the project's focus.

WP11: Erik Velldal (PhD) is a professor in the Language Technology Group at UiO/IFI and has been working with machine learning and NLP for ~25 years, publishing extensively in the top-tier venues of the field. He has been part of the NorLM effort at UiO, developing language models and benchmarking tools for Norwegian, and has also been instrumental in creating many of the available annotated datasets for

Norwegian NLP. He is currently the manager of the SANT project and also involved in other ongoing collaborations like the SFF Integreat, the SFIs MediaFutures and NorwAI, and the infrastructure project PSI. Velldal will be leading the WP on LLM benchmarking.

WP12: Michael A. Riegler (PhD) is an expert on building and researching complex AI systems specifically in sensitive areas such as medicine or child protection. AI systems in sensitive application areas often need a well-defined and conducted democratic process of stakeholder involvement. Riegler has led work packages in this context in several projects like for example the child avatar for police interview training of abused children. In addition, Riegler is researching different directions within the area of foundational models including training, finetuning and evaluation of these.

WP13: Klaus Johannsen (PhD) has an academic background in physics and mathematics, and a track record in management, computer science, interdisciplinary collaborations, including industry collaborations, machine learning, scientific computing, software engineering and project management. With this broad background he will be able oversee and coordinate the fine-tuning activities in WP13 in the best possible way to the overall benefit of the NORA.LLM project

WP14: Asbjørn Følstad (PhD) is Chief Researcher at SINTEF. His area of expertise is human-centred AI, with particular concern for user and stakeholder involvement in AI development and innovation. He has substantial project management expertise from European, Nordic and national research and innovation projects, and is currently involved in several European projects on human-centred and trustworthy AI. His background will be instrumental in achieving the WP14 objectives.

4.2 Plans for establishment/upgrade and operation

4.1.4 Workplan and deliverables

The NORA.LLM project will use an agile project management approach to ensure that we are adaptable to rapid change in the field. The project will have two phases: a developmental phase (years 1-3) and an operational phase (years 4-10). The main tasks that will be solved in the project are: Legal and copyright issues related to the use of data and open models; compiling Norwegian training datasets of sufficient size and quality; development of foundation models that are aligned with Norwegian values and cultures; enabling systematic benchmarking to assess both capabilities and limitations of the models; ensuring democratic and inclusive processes; optimization of models to solve specific tasks with a specific focus on health and education; developing business models for the post-project exploitation; updating existing high-performance computing in Norway. These tasks will be solved across the following 14 work packages: **WP1** Project management, **WP2** Communication and dissemination, **WP3** Legal framework for training LLMs and sharing training data, **WP4** Corpus creation, **WP5** Solution architecture, public Inference Infrastructure (hardware, software) and procurement, **WP6** Foundation model pre-training, **WP7** Foundation-model fine-tuning, **WP8** Chat optimization, **WP9** Retrieval, **WP10** Ethics and Alignment, **WP11** Benchmarking, **WP12** Democratic Process: An Inclusive AI Review and Policy-making Platform, **WP13** Domain-specific fine-tuning, and **WP14** Business and sustainability. **NOTE:** This is an extensive project with 14 WPs and more than 70 tasks and deliverables. The Gantt diagram is highly complex and is provided as an attachment.

4.1.5 Utilisation

The field is developing at a rapid pace and it is therefore difficult to predict the total usage of the infrastructure. However, the uptake of the technology is fast, as evident with ChatGPT – the fastest-growing consumer application in history, reaching 100 million monthly active users just two months after launch. In Norway, UiO recently released a speech bot that attracted 8000 users within a month. We assume a relatively rapid user-uptake of this infrastructure, both in academia, industry, and the public sector. Given the numbers above we would estimate 1.000 to 10.000 active users/month when the infrastructure is in the operational phase. From the research community the main Norwegian research centres have indicated their interest (Integreat, DSTrain; MediaFutures; Center for digital Narratives; Visual Intelligence, see attached letter of support). In addition, the infrastructure will be widely utilized by companies and the public sector. Many NORA partners, including NAV, the directorate of e-health, the Norwegian Cancer Register and Posten are actively pursuing this internally and will directly benefit from the infrastructure. The Health Directorate of Norway has expressed the need for LLMs in their project “*Enklere tilgang til informasjon*” (see attachment). The main outreach and work on user-uptake will be performed through the NORA network, including NORA.startup and the NORA research school.

4.1.6 Data management

NORA.LLM will provide support for FAIR data management. This will include access to training programs and consultation facilities for the researchers with priority for researchers from partner institutions. A

dedicated central storage facility, especially for data useful as training data, will be implemented using storage facilities provided by Sigma2. NORA.LLM will recommend all research projects to use existing Data Management Planning tools and promote data FAIRness using the upcoming UiO FAIR-system. NORA.LLM will be a “bring your own data” system, but it will also facilitate its own data lake to be utilized by all users, and for data sharing as granted by the data controllers. Metadata and datasets will be accessible to the general public once released by data controllers. All University partners and Sigma2 have long experience in supporting and training communities in data management.

4.1.7 Risk and mitigation measures

Tabell 1: Risk analysis and mitigation measures

Description of risk	Likelihood (Low/ Likely /Highly likely)	Consequence (Minimal/ Moderate/ Severe)	Work package(s) involved	Proposed risk mitigation measures
Personnel on temporary contracts leaving for permanent positions elsewhere.	Likely	Minimal	7, 8, 9, 10, 11	Allowing for re-announcement of shorter contracts.
Difficulty recruiting qualified personnel due to non-competitive salaries	Low	Moderate	7, 8, 9, 10, 11	Allowing for post-doctoral positions
Approaches to modelling considered state-of-the-art changes during the implementation phase	Highly likely	Minimal	6, 7, 8, 9, 10, 11, 13	Flexibly adopt most recent best practices. Agile management.
Access to sufficient data and quality due to GDPR and copyright	Likely	Severe	4, 6, 7, 11	Copyright: Agreements with media partners in collaboration with WP3 GDPR: implement pipelines for automatic anonymization and sanitization of texts
Models generate misinformation and/or harmful content	Likely	Severe	4, 6, 7, 8, 9, 10, 11, 13	Guidelines and pipelines for annotation of data and potential removal of harmful content.

4.2 Technical solutions

Minimizing risks at all stages is an important aspect when providing a critical national infrastructure, from design of the architecture and procurement, through deployment and operation. For the major procurements planned in this project we will leverage the long experience Sigma2 has on procurement for the national e-infrastructure for HPC and research data services. To reduce risks on the planned scale, it is necessary to base the significant part of investments on technologies with high Technology Readiness Level (TRL). At the same time, projections of future needs and emerging technologies.

The main investments in computing resources for the NORA.LLM infrastructure will be co-hosted with the national e-infrastructure implemented by Sigma2. Resources are carefully managed, upgraded and dimensioned according to documented needs in a stepwise manner. The national Resource Allocation Committee (RFK) decides, based on scientific merit and technical feasibility, how the available resources are allocated. Optimum technical solutions, enabling for the widest possible uptake and coverage of user’s needs in a national context, requires continuous efforts monitoring emerging technologies and building of competencies within the organization. Networking, including Norway’s participation in international e-infrastructure collaborations, is an important aspect of this.

In 2022, Sigma2 AS established a long-term contract with Lefdal Mine Data Center (LMD) for the provision of Housing services for the National e-infrastructure. The New storage system, NIRD is already installed and in production in these locations. Reference is made to chapter 2.2 - Ethical and environmental issues for more information regarding the locations for the e-infrastructure at LMD.

4.3 Cost- and funding plan

4.3.1 Cost center(s)/rental premises for the research infrastructure

The cost centers for NORA.LLM will be Sigma2, UiO and the National Library. The cost centers have equipment for HPC, data storage and data services. The partners will collaboratively be able to act as a

common facility and resell resources and competence services based on the RCN guideline for this (“leiestedsmodellene”). The sale of resources will follow the practice of the Sigma2 User Contribution model.

4.3.2 Costs and funding sources for establishment and operations

Tabell 2 – Cost of establishing and operating the infrastructure (numbers in NOK 1000)

Cost category	Establishing				Operational	Total
	Year 1	Year 2	Year 3	Total for years 1-3	Years 4-10	
Personnel (from RCN)	21 926	23 708	23 469	69 102	0	69 102
Personnel (in-kind)	10 476	13 893	14 290	38 659	20 961	59 620
Travel and workshops	800	800	800	2 400	0	2 400
Procurement	22 100	15 667	41 233	79 000	0	79 000
In-kind	11 175	14 093	14 990	40 259	21 661	61 920
Expected from RCN	44 826	40 175	65 502	150 502	0	150 502
Total	56 001	54 268	80 492	190 762	21 661	212 423

Footnote: In-kind contributions include personnel costs, computing, and hardware.

4.3.2.1 Funding for the period from year 4 to through 10.

Sigma2, UiO-IFI and NORA will together contribute with the following in-kind person-months (PMs) during the operational phase: 42.4 PMs in year 4; 19.7 PMs in year 5; 16.7 PMs in years 6-9, and 15.7 PMs in year 10. Basic operational costs (power, cooling and footprint) are included as part of the operational costs of the national e-infrastructure provided by Sigma2 for the major investment in equipment in year 3. By this arrangement, the operational costs will be covered by projects using the infrastructure through research funders like the NRC and the EU, or through the Sigma2 cost-model for users with own funding. To some extent (less than 20% of total capacity) the NORA.LLM infrastructure services could also be “sold” to commercial users.

4.3.3 Use of existing national infrastructures.

NORA.LLM will use existing national infrastructures (Sigma2). The cost associated with this will be partially covered according to the RCN mandated User Contribution Model. This includes the use of the LUMI system.

4.3.4 Down-scaled budget

To cut budget, the models can be downsized (not training the larger models), or we can build the Norwegian model around already existing models. Fine tuning of the models into specific domains can in principle also be cut down. We can also buy less physical hardware and reduce travels and workshops:

- 50% reduction in capital expenditure from 80 MNOK to 40 MNOK
- 50% reduction in travel and workshop from 2,4 MNOK to 1,2 MNOK
- 20% reduction in personnel costs from 129 MNOK to 103 MNOK

This will lead to a total budget reduction of 67 MNOK, from 212 MNOK to 145 MNOK.

Effect of reduced budget on the work plan Reduced budget will not impact the timelines of the work plan, but with less personell and less available hardware we will have not be able to build the larger models of the infrastructure from scratch.

Possibility of incremental build-up of infrastructure In principle, the hardware and models can be incrementally upgraded. The original plan is to train models from scratch (the estimated compute for a functional 10 billion parameter model is 100K GPU hours). The incremental alternative is to fine-tune existing pre-trained models and/or lower the amount of GPU hours allocated for model training (from 100K to ca. 50K GPU hours per model). Hardware could be acquired according to reduced needs. The result would be a reduction in CapEx of approx. 50% (-20 M in model training and -20 M in hardware).

Impact of reduced budget Fine-tuning a pre-trained model on a large Norwegian corpus is still essential, and it will likely require significant computing resources. With pre-trained models, there will be other issues, like the use of non-Norwegian tokenizers, which in many instances cannot encode Norwegian-specific characters ("æ", "ø" and "å"). Equally important is the inability to modify the hyperparameters, experiment with architectures, etc, when using pre-trained models. This will also have implications for the transparency of the models and the potential downstream research. Reducing the GPU hours allocated to training is entirely possible, however, this will have a significant impact on the final performance of the models or the model sizes.

Acquiring GPUs based on needs as they arise in time will ensure that only the latest GPUs are procured, however, unless funds are specifically allocated to hardware, we might end up in a situation with insufficient computing resources to operate the infrastructure. Additionally, smaller and/or incremental investments in hardware will result in a significantly poorer economy of scale, as each installment incurs costs associated with procurement and deployment for operation.

A way to mitigate reduced budget for hardware and/or rented computing resources is to apply for computing time for training etc., i.e. from EuroHPC. Actual hardware investments could in principle then be limited to what is needed for inference and associated core services. The major downside with this would clearly be the uncertainty and lack of predictability. Sigma2 and/or NAIC resources could be used to some extent in the same manner, but the contribution model (Norwegian leiestedsmodell) would still require a budget according to the policies established for use of research infrastructures by the RCN (reflecting that there would be a cost within Norway).

5 References

- [1] gov.uk, "Tech entrepreneur Ian Hogarth to lead UK's AI Foundation Model Taskforce," 2023. [Online]. Available: <https://www.gov.uk/government/news/tech-entrepreneur-ian-hogarth-to-lead-uks-ai-foundation-model-taskforce>. [Accessed 2023].
- [2] Kommunal- og distriktsdepartementet, "The National Strategy for Artificial Intelligence," 14 01 2020. [Online]. Available: <https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=2>. [Accessed 2023].
- [3] Kunnskapsdepartementet, "Langtidsplan for forskning og høyere utdanning 2023–2032," [Online]. Available: <https://www.regjeringen.no/no/dokumenter/meld.-st.-5-20222023/id2931400/?ch=1>.
- [4] Research Council of Norway, "Norwegian Roadmap for Research Infrastructure 2023," 04 09 2023. [Online]. Available: <https://www.forskningsradet.no/en/apply-for-funding/funding-from-the-research-council/infrastruktur/veikart-for-forskningsinfrastruktur/strategic-basis/>. [Accessed 2023].
- [5] NorwAI, "NorwAI Research & Innovation," [Online]. Available: <https://www.ntnu.edu/norwai/research>. [Accessed 2023].
- [6] AI Sweden, "GPT-SW3," [Online]. Available: <https://www.ai.se/en/node/81535/gpt-sw3>. [Accessed 2023].
- [7] LUMI, "The waste energy of LUMI supercomputer," 05 2020. [Online]. Available: <https://www.lumi-supercomputer.eu/the-waste-energy-of-lumi-supercomputer-produces-20-percent-of-the-district-heat-of-kajaani-csc-and-loiste-lampo-have-signed-an-agreement/>. [Accessed 2023].
- [8] UiB, "Center for Digital Narrative," [Online]. Available: <https://www.uib.no/en/cdn>. [Accessed 2023].
- [9] Integreat, "Integreat," [Online]. Available: <https://www.integreat.no/>. [Accessed 2023].
- [10] Visual intelligence, "Visual intelligence," [Online]. Available: <https://www.visual-intelligence.no/>. [Accessed 2023].
- [11] Media Future, "Media Future," [Online]. Available: <https://mediafutures.no/>. [Accessed 2023].
- [12] titan.uio.no, "MN-fakultetet skal ansette 36 nye forskere, mange av disse innen maskinlæring og KI," [Online]. Available: <https://www.titan.uio.no/teknologi/2023/mn-fakultetet-ved-skal-ansette-36-forskere-innen-m.html>. [Accessed 2023].
- [13] UiB, "Introduction to LEAD AI," [Online]. Available: <https://www.uib.no/en/ii/164661/introduction-lead-ai>. [Accessed 2023].
- [14] R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong and P. Liang, "The Foundation Model Transparency Index," 19 10 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2310.12941>. [Accessed 2023].
- [15] Universitet i Oslo, "UiOs Veikart for forskningsinfrastruktur," UiO, Oslo, 2022.