Helsedirektoratet

Report on large language models in Norwegian health and care services: Risks and adaptations to Norwegian conditions

Først publisert: 29. juli 2025
Last revised date: 29. juli 2025

Innhold

Foreword

The Ministry of Health and Care Services has tasked the Norwegian Directorate of Health with developing a joint AI plan for the health and care services, in collaboration with the Norwegian Medicines Agency (NOMA), the Norwegian Institute of Public Health (NIPH), the Norwegian Radiation and Nuclear Safety Authority (DSA), the Norwegian Board of Health Supervision, the regional health authorities, and the Norwegian Association of Local and Regional Authorities (KS). As part of the work on this plan, the Norwegian Directorate of Health has been tasked with conducting an assessment "of what risks large language models introduce and identify measures to ensure that the health and care services have access to language model(s) that are adapted to Norwegian conditions." This report addresses that mandate.

The work has been carried out through interviews with experts and literature studies of relevant research articles and reports related to research on, development of, and use of large language models in Norway. The following persons and institutions have been involved in the work: Benjamin Kille (NTNU/NorwAI), Svein Arne Brygfjeld (National Library of Norway), Geir Thore Berge and Tor Oddbjørn Tveit (Sørlandet Hospital), Troels Sune Mønsted (University of Oslo), Christian Autenried (Helse Vest ICT), Damoun Nassehi (General Practitioner/UiB), Michael A. Riegler (Simula and OsloMet), Jens Osberg (Norwegian Digitalisation Agency), Karl Øyvind Mikalsen (Center for Patient-Centered Artificial Intelligence (SPKI)), Kristine Eide, Matias Jentoft and Bjarte Engelsland (Language Council). We wish to thank them for willingly participating in interviews throughout the project. We extend special thanks to Michael Riegler for close follow-up, advisory work, and quality assurance of texts.

Chapter 3 on risks in the use of large language models in health and care services and Chapter 4 on adaptations to the Norwegian health and care services provide the knowledge base and background for recommended measures.

The report reflects knowledge at the beginning of 2025. AI is a rapidly developing field and new legislation from the EU will affect the use of AI in Norway. This report recommends measures that should be initiated now to lay the foundation for safe and responsible use of AI going forward.

Prioritization of recommended measures will be done in dialogue with the Ministry of Health and Care Services, the AI Council, and the leadership of the Norwegian Directorate of Health. Prioritized measures will be incorporated into the Joint AI Plan for Health and Care Services on an ongoing basis. A majority of the measures will extend beyond the planning period for the joint AI plan, which lasts until the end of 2025.

Oslo, April 1, 2025

Executive Summary

The potential and expectations for artificial intelligence (AI) are high in health and care services, which is reflected in the national digitalization strategy and the National Health and Coordination Plan [1][2].The latter points out that AI can contribute to faster and more precise diagnostics, better decision support for healthcare personnel, simplified logistics, automation of administrative tasks, and enable citizens to better monitor their own health. AI will also make a significant contribution to sustainable development of our shared health service.

Development and use of artificial intelligence (AI) and especially large language models has seen significant growth in recent years. Use of large language models has the potential to improve both efficiency and quality in health and care services, for example through automatic text production, structuring of free text in patient records, machine-assisted health professional coding, plain language assistance, translation, speech recognition, knowledge support, health research, sorting and triaging patient inquiries, clinical chatbots (questions and answers), citizen-directed chatbots, health education, and decision support in treatment [3].

On the other hand, there are significant risks associated with adopting generative AI. This report describes important risks that the Norwegian health and care services must consider to be able to use AI systems based on large language models in a responsible manner. The risks relate both to inherent risks associated with this type of AI models and to their use in health and care services.

Currently, language models are best suited for linguistic and administrative tasks. Quality assurance is important regardless of use. AI systems with language models that are to be used to provide healthcare are most likely medical devices, and thus regulated through the act on medical devices.

Adaptation to Norwegian conditions is one way to reduce risks associated with the use of large language models in health and care services. The report describes what such adaptation entails, how adaptations can be made, what is needed, and who can make adaptations.

Finally, the report recommends measures that can reduce risks and facilitate use that is adapted to Norwegian conditions in health and care services:

- Establish a quality framework for large language models for Norwegian health and care services
- Build and share competence on development and use of large language models
- Establish common data foundation
- Investigate the need for infrastructure with computing power adapted to the health sector's needs
- Investigate governance structure for large language models in the health and care sector

Most measures will need to be solved in collaboration between a number of actors, such as the Norwegian Directorate of Health, Norwegian Health Network (NHN), the Research Council of Norway, the National Library of Norway, the higher education sector, regional health authorities, regional ICT companies, and KS. It is proposed that the Norwegian Directorate of Health take particular responsibility for establishing a quality framework.

[1] https://www.regjeringen.no/en/dokumenter/the-digital-norway-of-the-future/id3054645/

[2] https://www.regjeringen.no/no/dokumenter/meld.-st.-9-20232024/id3027594/

[3] The Norwegian Directorate of Health has, among other things, written about opportunities with large language models in this report:
https://www.helsedirektoratet.no/rapporter/store-sprakmodellar-i-helse-og-omsorgstenesta

Introduction

The potential and expectations for large language models in health and care services are high, while the risks are significant. It is crucial that AI systems with large language models used in health and care services are trustworthy, and the report has identified five measures that will contribute to this.

The American National Institute for Standardization and Technology (NIST) characterizes trustworthy AI as: Valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed [4]. The European Union published ethical guidelines in 2019 describing that a trustworthy AI system has three components that should be fulfilled throughout the system's lifecycle: (1) it should be lawful, complying with all applicable laws and regulations, (2) it should be ethical, ensuring adherence to ethical principles and values, and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm [5].

Objectives

The objective of this report is threefold:

1. To describe and clarify risks associated with the use of generative AI and large language models in health and care services
2. To describe how large language models can be adapted to the Norwegian health and care services, as one of several possible risk-reducing measures
3. To propose measures that will contribute to making large language models well adapted to conditions in the Norwegian health and care sector

The report is to be considered as a basis for discussions about prioritization of and responsibility for relevant measures that will contribute to this.

Target audience

The main target audience for the report is the Ministry of Health and Care Services, the AI Advisory board for the health and care sector, the health and care sector, collaboration partners (such as the National Library of Norway and R&D institutions), and other authorities relevant for initiating the recommended measures.

The report will also be of interest to those working with generative AI and health.

Scope

This report concerns generative artificial intelligence, mainly large language models. Large multimodal models are also briefly covered.

In May 2024, the Norwegian Directorate of Health published the knowledge base "Large language models in health and care services" [6]. The document describes how such models can contribute to improving health services, potential benefits and challenges, and how large language models can be adapted for use in health and care services. This report does not cover the opportunity space and elaborates on challenges and how language models can be adapted to health services.

In January 2025, the Norwegian Directorate of Health published "Report on quality assurance: Use of artificial intelligence in health and care services" [7]. The report covers, among other things, risk management and quality assurance related to implementation and use of AI in general. This report goes deeper into risks and quality assurance related to generative AI models and especially large language models.

The Norwegian Directorate of Health develops AI services that use large language models, including in Helsesvar and "Enklere tilgang til informasjon" (ETI). Privacy issues related to these projects are addressed in the Data Protection Authority's regulatory sandbox and subsequent report (which was not published when this report was completed) [8]. This report also addresses privacy-related risks but does not go into detail on specific use cases.

Method

The work on the report has been done through a combination of interviews and literature studies of Norwegian and international research articles and relevant reports.

We have interviewed professionals with experience or deep knowledge related to research on, development, and use of large language models in Norway. The interviews were conducted from October 2024 to February 2025. The following institutions have been involved in the work: NTNU, NorwAI, National Library of Norway, Sørlandet Hospital, University of Oslo, Helse Vest ICT, a general practice clinic, University of Bergen, Simula Center, OsloMet, Norwegian Digitalisation Agency, Center for Patient-centered Artificial Intelligence (SPKI), and the Language Council of Norway.

Key terms

The definition of an AI system according to the AI Act, and as we use the term in this report is: "...a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" [9].

An AI system can take various forms: as a standalone application on a PC or mobile application, a web service, part of a professional system, or a health record system.

An AI system is a medical device under the Act on medical devices if it is intended to be used on humans for the purpose of contributing to diagnosis, prevention, monitoring, prediction, prognosis, treatment, or alleviation of disease [10][11].

An AI system can consist of one or more AI models, in addition to improvements such as knowledge grounding.

An algorithm is a recipe that tells how something is done, and simply explained consists of a set of instructions that transform input data into output data. A machine learning algorithm is the recipe the system uses to build an AI model of reality, based on the data that is fed in (the training data). This model can then be used to make new decisions about new incoming data [12].

Foundation models are large neural networks trained on large general datasets that may consist of text, images, audio, etc. Such models can form the basis for a range of different AI solutions [13].

Generative AI encompasses solutions that primarily generate or produce new material, such as text or images [14]. Generative AI models differ from non-generative AI models in the following ways:

Generative AI models create new content, such as text, images, or music, based on training data. For example, a generative model can create new images of cats based on previous cat images.

Non-generative, including classification models, AI models distinguish between different classes or predict probabilities for specific outcomes. They focus on finding boundaries between categories in the data. A classification model can determine whether a given image contains a cat or not.

Large language models (LLM) are generative AI models. Large language models, like the well-known GPT models, are a form of foundation models trained on enormous amounts of text to predict what comes after a word or syllable in a given context. Language models do not store the text they are trained on. They "know" nothing about the world, and they do not browse websites to find facts. They only know language. Nevertheless, it is easy to believe that the models think, or know something, because they are so good at writing text that we humans experience as meaningful [15].

More recently, large language models have been extended to handle other modalities, such as images, audio, and video. These are called multimodal AI models, but the term "LLM" is still often used. This is due to several factors:

The core of the model is linguistic: Even though they can process multiple modalities, the text interface is still the primary way users interact with the model.

Underlying technology: The models build on architectures developed for language processing (like transformers), and much of the training occurs on large text datasets.

General professional terminology: Within the AI field, the term "LLM" is still used broadly, as most multimodal models spring from LLMs that have been extended to interpret other data types.

In this report, we use the same definition of risk as in the AI Act: "Risk means the combination of the probability of an occurrence of harm and the severity of that harm" [16].

Other terms used in the report are described in the AI fact sheet "AI Terms" [17].

[4] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf p.12

[5] https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en

[6] https://www.helsedirektoratet.no/rapporter/store-sprakmodellar-i-helse-og-omsorgstenesta

[7] https://www.helsedirektoratet.no/rapporter/report-on-quality-assurance-use-of-artificial-intelligence-in-health-and

[8] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/pagaende-prosjekter2/helsedi

[9] Aticle 3 in the AI Act https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

[10] https://lovdata.no/dokument/NL/lov/1995-01-12-6

[11] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02017R0745-20200424

[12] https://media.wpd.digital/teknologiradet/uploads/2022/02/Kunstig-intelligens-i-klinikken.pdf p.16

[13] https://www.regjeringen.no/contentassets/c499c3b6c93740bd989c43d886f65924/en-gb/pdfs/digitaliseringsstrate p.69

[14] https://www.regjeringen.no/contentassets/c499c3b6c93740bd989c43d886f65924/en-gb/pdfs/digitaliseringsstrate p.69

[15] https://www.regjeringen.no/contentassets/c499c3b6c93740bd989c43d886f65924/en-gb/pdfs/digitaliseringsstrate p.69

[16] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689#cpt_I Article 3(2)

[17] Fact sheet "AI Terms"

Risks associated with using large language models in health and care services

Use of large language models has great potential to improve both efficiency and quality in health and care services, for example through structuring free text in patient records, machine-assisted health professional coding, decision support in treatment, knowledge support, predictions, sorting and triaging patient inquiries, automatic text production, clinical chatbots (questions and answers), citizen-facing chatbots (questions and answers), plain language assistance, speech recognition, health education, health research, translation, and logistics [18].

At the same time, language models have some inherent challenges that can affect both patient safety and the quality of health services, which in turn can affect overall trust in health and care services. Use of large language models is in an early phase, and there are still uncertainties related to benefits and gains in the short and long term. Researchers at the Simula Institute point out that "artificial intelligence (AI), and especially generative AI, is a collection of seductive technologies that, if one lacks good technological insight, can mislead decision-makers at all levels in organizations to make uninformed assessments" [19].

There are also uncertainties related to which areas large language models can be used in health and care services in a responsible and appropriate manner. Currently, they are best suited for linguistic and administrative tasks with low risk. Quality assurance is important regardless of use.

AI systems with large language models intended for providing healthcare have higher risk. They are most likely medical devices and thus regulated through the act on medical devices [20]. The purpose of this law is to prevent harmful effects, accidents, and incidents, and ensure that medical equipment is tested and used in a professionally and ethically sound manner. As of March 25, we are not aware of AI systems with generative language models that are CE-marked after conformity assessment performed by a notified body.

The AI Act has entered into force in the EU with the goal of establishing trust in AI use in the EU while facilitating innovation. The regulation has a risk-based approach, where requirements for an AI system are determined by the risk level of the intended use.

Furthermore, both users and developers of AI systems to be used in health and care services in Norway must comply with both general and sector-specific regulations. For example, sector-specific regulations such as Norwegian health legislation, the EU's General Data Protection Regulation, copyright law, and equality and anti-discrimination law apply.

This chapter describes both underlying challenges with large language models and risks related to use in health and care services. The included risks are not exhaustive.

[18] The Norwegian Directorate of Health has, among other things, written about opportunities with large language models in this report:
https://www.helsedirektoratet.no/rapporter/store-sprakmodellar-i-helse-og-omsorgstenesta

[19] https://www.digi.no/artikler/debatt-ki-feberen-nar-middelet-blir-malet/554186

[20] https://lovdata.no/dokument/NL/lov/1995-01-12-6

Risks in large language models

Large language models are very complex, and the generated responses are based on statistical patterns in training data rather than actual understanding of content. Errors, biases, and lack of transparency can have consequences for patient safety and quality of health and care services. This chapter describes some of the underlying causes that come from training and characteristics of large language models.

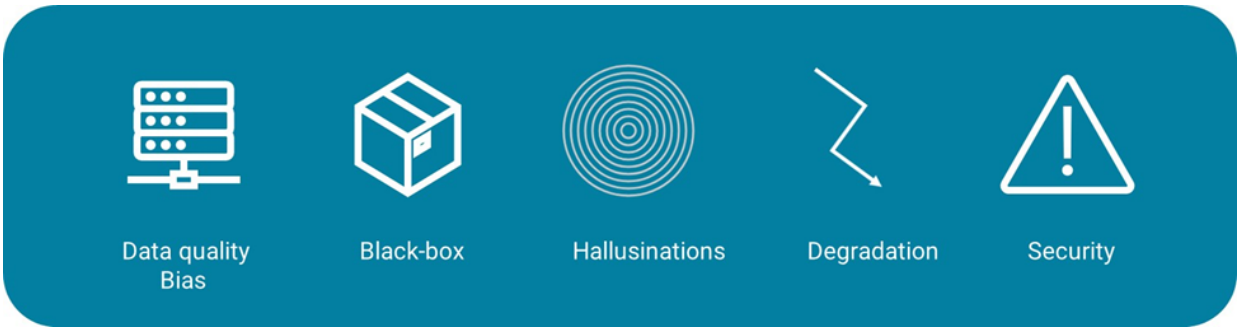

Figure 1: Underlying uncertainties and characteristics of large language models that contribute to risk when used in health and care services include data quality and bias in the data foundation, "black box" problems, hallucination, model degradation over time, and vulnerabilities in cybersecurity and information security.

Poor data quality and bias

Training large language models requires enormous amounts of data collected from sources such as books, journals, websites, and social media. Good sources can also be patient records and medical images.

The quality of the data will affect the performance of the models. Some factors that affect data quality are:

- *Incomplete or deficient data.* Dropout of enrolled patients in clinical studies can lead to incomplete data. Lack of registered information leads to incomplete and deficient data [21]. This can be because the data has not been collected systematically. Patient records may also lack information due to varying practices in data registration [22]. Another reason may be that privacy considerations can result in omission of data.
- Misinformation refers to information that is false or misleading [23]. Disinformation is a systematic and deliberate form of misinformation [24]. Small amounts of misinformation in training data have been shown to lead to misleading responses from medical AI models [25]. Large language models may be trained on data from the internet and/or third-party data sources of varying quality. These sources may contain misinformation that can be difficult to detect.

AI models can perpetuate and amplify biases found in training data, as well as contribute to reinforcing biases that affect human assessments and judgment [26][27]. Below are some examples of types of biases in data that are important to be aware of, especially in the health field:

- *Overrepresentation of historical data* compared to newer data makes the responses from models not updated according to current practices and research. In medicine and health, new knowledge should be weighted heavily compared to historical practice.
- *Selection bias* in, for example, publication of scientific articles. Positive findings and data are published more often than negative findings [28]. Clinical studies with significant results are reported more often than negative studies [29]. This means that the texts and publications used to train large language models are systematically biased toward positive findings and studies that confirm

hypotheses, while negative results are underrepresented. When reporting intervention studies, for example, benefits are often emphasized more than disadvantages.

- Lack of local context. Training data for large, general language models has a strong overrepresentation of text and information in English. In some cases, models are trained only on English text. Norwegian content from the internet will in any case constitute a very small part of training data [30][31]. Watson Oncology is an example of how lack of local context can affect a model's performance [32]. When the model was tested in Denmark, it had considerably worse performance than previously reported. Danish clinicians pointed to both major differences between Danish and American treatment guidelines for cancer, as well as poorer quality of data from clinical studies conducted in the USA as possible causes.
- Skewed or incomplete data for certain groups and minorities. Not all groups and minorities are equally represented in the data used to train large language models [33]. Historically, for example, women have been underrepresented in clinical studies [34]. People with disabilities or minority backgrounds may also be underrepresented in clinical studies. This may be due to deliberate exclusion from studies, or practical barriers such as transport, economy, or other obstacles to participation [35][36.] This is also relevant internally in Norway where there may be less available data from minorities and the Sami indigenous population [37].

Lack of transparency, explainability, and interpretability ("black-box" problem)

Large language models can function as "black boxes," where neither users nor developers can fully explain how a specific response (output) from an AI model is generated. Lack of transparency, explainability, and interpretability can contribute to this "black box" problem [38]. These are distinct characteristics that build on each other.

Transparency refers to *what* happened in the model. Indicators that contribute to transparency include available information about:

- Data used for training
- Source code
- Training methods
- Evaluation
- Performance measures

The Foundation Model Transparency Index is based on 100 different indicators of transparency related to training, the model itself, or use of the model [39]. For example, Mistral 7B, which can be used by third parties without restrictions, has an index of 55%, meaning that there is no information available for 45 of the 100 indicators (including data sources) [40].

Explainability refers to how an AI system works and what mechanisms govern the decisions it makes. With many millions or billions of parameters and complex neural networks, large language models can operate as black boxes, even with openness around both data and source code [41].

Interpretability refers to *why* the model arrived at a response [42]. This can be showing users what instructions were given to arrive at the answer, which current AI systems rarely show. A method that contributes to better interpretability is reasoning (see AI fact sheet Intelligence enhancement and control) [43].

Large language models operate probabilistically and generate the most likely response based on training data, but often without clear understanding of when responses are correct. Lack of explainability and

interpretability can, for example, make it difficult for clinicians to assess whether recommendations from language models are reliable, because the underlying mechanisms and data are opaque [44].

Methods for grounding models in facts, making reasoning visible, or showing which parameters are weighted (see AI fact sheet Intelligence enhancement and control [45]) can make it easier for healthcare personnel to assess the quality of the generated response [46] At the same time, use of grounding methods can reduce attention to the model's limitations that still require critical evaluation of outputs [47].

The lack of explainability and interpretability also makes it difficult to detect when a model generates incorrect or fabricated responses, also known as hallucinations.

Hallucination

Because language models do not have built-in understanding of knowledge and truth, but operate based on probability calculations, they can generate incorrect, incomplete, inaccurate, or misleading information with great confidence. This phenomenon, known as hallucination, means that the model can present information that looks credible but is either wrong, taken out of context, or has no basis in training data.

The probability of hallucination increases if training data has low quality or lacks representation, for example if important medical perspectives or updated guidelines are missing. Even models trained on high-quality data can hallucinate, as this is an inherent limitation of large language models [48].

Hallucination can be divided into two categories:

- Fact-hallucination, where the model generates false or fabricated claims
- Context-hallucination, where information is misinterpreted or placed in an incorrect context

In addition, omission of critical information can pose a risk, particularly when strict accuracy is required such as in health contexts.

Temperature setting (see fact box about temperature), post-training, and methods for knowledge grounding can reduce the occurrence of hallucination (see chapter 4.2).

Temperature (degree of variation)

In the field of AI, temperature is a parameter that controls how random the generated text should be. Temperature can be adjusted in the selection phase, where the model chooses the next word based on calculated probabilities. Low temperature (for example 0.1-0.3) makes the model more often choose the most probable words, giving more predictable and conservative responses. High temperature (for example 0.8-1.0) increases the chance that it chooses more unexpected or less probable words, giving more creative and varied responses. For medical applications, choosing lower temperature would ensure more precise and consistent responses.

Example (generated by ChatGPT-4o): Question: "*What are common causes of headache?*"

- *Low temperature* (0.2) gives a concise, accurate, and objective response: "*Common causes of headache include stress, dehydration, muscle tension, sleep deprivation, and migraine. Some medical conditions, such as high blood pressure or sinusitis, can also cause headache*."
- *High temperature (0.8)* gives a more varied, informal, and descriptive response*: "Headache can be caused by everything from stress and dehydration to poor sleep or tense neck muscles. Some get headaches from strong smells or loud sounds, while others may experience it as a symptom of*

underlying diseases such as sinusitis or migraine. In rare cases, headache can be a sign of more serious conditions."

Degradation of model performance over time

Large language models are dynamic and influenced by both external and internal factors over time. Three challenges related to degradation of model performance are model drift, catastrophic forgetting, and model collapse. These phenomena can reduce the model's ability to provide accurate, relevant, and reliable responses, which is critical in health contexts.

Model drift describes how a model's performance can weaken over time, either because it encounters new data that differs from the training foundation, or because the relationship between input and output changes [49]. Model drift can lead to AI systems giving outdated or incorrect recommendations or information. This is relevant in the health field, where patient populations, treatment methods, and medical guidelines constantly evolve. Some underlying causes are:

- *Data drift* which occurs when the distribution of input data the model encounters in daily use differs from the data used to train and optimize the model. For example, a model developed for a younger patient population may give less precise recommendations if used instead on an older patient group.
- *Concept drift* which happens when the relationship between input and output changes over time. For example, if a new treatment or lifestyle change proves to reduce the relationship between age and diabetes risk, the model will give outdated or misleading recommendations because it still relies on the old relationship.
- *Outdated information.* It is resource-intensive to train large language models, and they will not always be trained on the newest available information. This can lead to models giving outdated responses that affect model performance in medicine and health [50][51].

Catastrophic forgetting (catastrophic interference) occurs when a model loses or degrades previously learned knowledge as a result of training on new data. For example, a large language model may lose precision for medical diagnostics if it is post-trained on general text.

Catastrophic Forgetting

Causes of catastrophic forgetting include:

1. updating all weights in the model during re-training
2. sequential learning where old data is no longer available
3. overfitting to new data, displacing previous learning

There are many strategies for continual learning over time that can counteract catastrophic forgetting. These include *selective weight freezing* to preserve critical knowledge, *experience replay* techniques that mix historical data with new training data, or architecture-based solutions such as progressive neural networks where new tasks are handled in separate layers. These methods can be implemented individually or in combination to achieve more robust learning over time.

Source: https://link.springer.com/article/10.1007/s11063-024-11709-7

Model collapse is a long-term challenge that can occur when an increasingly large proportion of training data is generated by language models rather than humans. This can lead to models learning from previous models instead of from genuine human knowledge, which gradually weakens the quality and diversity of training data [52]. Model collapse is not just about synthetic data, but also about loss of model complexity

and variation. When models learn from each other, subtle patterns, nuances, and extreme values in the data can disappear. If this leads to gradual degradation of the model's ability to capture complex relationships and produce varied content, it can result in irreversible errors or defects. This risk is not acute but may become a challenge in the future if large amounts of future data are synthetically generated.

Vulnerabilities in cybersecurity and information security

Cybersecurity must be part the supply chain assessment of a language model. Traditional cybersecurity measures such as access control, logging, and testing are still relevant but must be adapted to the unique vulnerabilities of language models. Risk assessments should additionally include the model's access levels, data handling, and protection against input manipulation to ensure safe solutions in the health and care sector.

Large language models differ from traditional software in their complexity and lack of transparency (see under "Lack of transsparancy, explainability, and interpretability ("black-box" problem) in this chapter), which creates particular cybersecurity challenges. Unlike traditional software, where source code can be reviewed for security assessment, much of the functionality of language models is based on enormous datasets that can be difficult to verify (see under "Poor data quality and bias" in this chapter). Often the foundation model itself is a "black box," and the supply chain is complex with various actors for model development, data, and software. This makes it challenging to detect vulnerabilities related to, for example, poisoned training data or backdoors in the model.

Language models have broad areas of use and can be used in unforeseen ways. If they have access to documents or systems, they can inadvertently expose sensitive information. An example is Copilot, which has access to all data the user has access to, which places great demands on control routines [53]. In the health sector, where integrity and confidentiality are crucial, this is a particular challenge.

A known risk is prompt injection, where attackers manipulate the model's responses using specially designed queries to the system. This way, an attacker can, for example, bypass security mechanisms or reveal training data. Indirect prompt injection, where malicious input comes from external documents or websites, is particularly concerning as it can affect the response without the user becoming suspicious.

Large language models require significant resources, making them vulnerable to Distributed Denial of Service attacks (DDoS) through many or resource-intensive queries.

The Open Worldwide Application Security Project (OWASP) [54] and MITRE [55] point to a range of specific security risks with language models and update them regularly in pace with technological development.

[21] https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-tr

[22] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2776905

[23] https://tenk.faktisk.no/ordbok Norwegian term: Feilinformasjon

[24] https://tenk.faktisk.no/ordbok Norwegian term: Desinformasjon

[25] https://www.nature.com/articles/s41591-024-03445-1

[26] https://arxiv.org/abs/2201.11706

[27]https://www.nature.com/articles/s41562-024-02077-2

[28] https://ebm.bmj.com/content/24/2/53.long

[29] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114023

[30] https://arxiv.org/abs/2101.00027

[31] URLs for the .no domain constituted 0.3% of all URLs in the web-crawl for Common Crawl February 2025: https://commoncrawl.github.io/cc-crawl-statistics/plots/tld/latestcrawl.html

[32] https://link.springer.com/article/10.1007/s00146-020-00945-9

[33] https://www.who.int/publications/i/item/9789240084759 p.10

[34] https://cordis.europa.eu/article/id/27270-exclusion-from-clinical-trials-harming-womens-health

[35] https://gh.bmj.com/content/8/11/e013473

[36] https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.23157

[37] https://uit.no/research/sshf-no?p_document_id=674134&Baseurl=/research/#region_752662 The Centre for Sami Health Research (SSHF) was established in 2001 because of lacking knowledge about health and living conditions of the Sami population.

[38] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf p.16

[19] Center for Research on Foundation Models (CRFM) at Stanford: https://crfm.stanford.edu/fmti/paper.pdf

[40] https://crfm.stanford.edu/fmti/May-2024/index.html 2024

[41] https://www.ibm.com/think/topics/black-box-ai

[42] https://www.ibm.com/think/topics/interpretability

[43] https://www.helsedirektoratet.no/digitalisering-og-e-helse/kunstig-intelligens/ki-faktaark

[44] https://ai.nejm.org/doi/pdf/10.1056/AIra2400380

[45] https://www.helsedirektoratet.no/digitalisering-og-e-helse/kunstig-intelligens/ki-faktaark

[46] https://ai.nejm.org/doi/pdf/10.1056/AIra2400380

[47] ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf

[48] https://arxiv.org/abs/2401.11817

[49] https://www.ibm.com/think/topics/model-drift

[50] https://www.who.int/publications/i/item/9789240084759

[51] https://arxiv.org/abs/2303.08774

[52] https://www.nature.com/articles/s41586-024-07566-y

[53] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rap

[54] https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[55] https://atlas.mitre.org/

Risks associated with use in health and care services

Risk is closely linked to areas of use. The consequences of the inherent challenges with large language models, such as hallucinations, bias, and lack of transparency (chapter 3.1), can vary from insignificant to critical, depending on how and where the models are used. Some challenges arise in the use situation while others emerge over time and with increased use.

Some risk areas can have consequences for patient safety and quality in health services (Figure 2): the dialogue between human and machine, dependency that can arise with extensive use of AI tools, and privacy in handling sensitive data. Other risks are more overarching and relate to complexity in and interaction between various regulations, unintended illegal use of large language models, discrimination against individuals or groups based on biases in the models, environment and sustainability, and where or with whom responsibility lies if errors occur.

Figure 2: Risks related to the use of large language models in health and care services. Risks include inadequate competence in dialogue with language models, reduced professional knowledge and skills, reduced privacy, complex regulations, unintended illegal use, discrimination, environment and sustainability, scientific evidence, and responsibility in case of harm.

Through the work on the report, it has become clear that there are particular uncertainties associated with development and use of generative AI with high risk, such as in diagnostics and treatment. Among other things, WHO questions whether large language models will be able to achieve sufficient accuracy to justify costs and resources associated with development, and safe and effective implementation of such tools in given areas within health care [56].

The American authorities (Food and Drug Administration (FDA)) have assessed medical equipment with generative artificial intelligence (AI). They point to several challenges and risks such as hallucination, lack of transparency, continuous learning, challenges with scientific documentation, and the need for new evaluation methods (see fact box about FDA report below) [57]. We are not aware of any similar assessments published by the EU.

Assessments related to medical equipment with generative AI conducted by FDA

FDA has assessed medical equipment that uses generative artificial intelligence (AI) and points to several challenges and risks:

- *Hallucinations*: Generative AI can produce incorrect content ("hallucinations"). For example, equipment designed to summarize a conversation between a patient and healthcare personnel may inadvertently generate a false diagnosis that was never discussed.
- *Lack of transparency:* Equipment using foundation models developed by third parties often has limited access to information about the models' architecture, training methods, and datasets. This can make it difficult for manufacturers to ensure quality and safety.
- *Continuous learning:* Generative models can either be static or change continuously ("continuous learning"). Continuous learning is associated with uncertainty about the model's performance over time, and as of November 2024, FDA had not approved medical equipment using continuous learning.
- *Challenges with scientific documentation:* It can be difficult to determine what type of valid scientific documentation ("evidence") FDA should request to be able to assess the equipment's safety and effectiveness throughout the lifecycle.
- *Challenges with FDA's classification system:* Generative AI can introduce new or different risks that challenge FDA's current classification system for medical equipment. How equipment is classified affects what regulatory measures are necessary to ensure the equipment becomes safe and effective.

FDA also highlights challenges related to evaluation and testing of generative models:

- *Pre-market evaluation*: Large language models have complex parameters and can give different responses based on small changes in wording or prompts. It is not possible to test all possible prompts before launch. Moreover, lack of transparency and the potential for unforeseen responses can make it particularly difficult to evaluate such systems before they come to market. They (FDA) point to the importance of plans and methods for monitoring after the equipment is put into use (postmarket monitoring).
- *Need for new evaluation methods*: Current methods for quantitative performance evaluation may be insufficient to ensure safe use of generative AI. New, qualitative evaluation methods can help characterize/describe the model's autonomy, transparency, and explainability. Which evaluation methods are required will vary based on the product's specific use area and design.

FDA recommends manufacturers of medical equipment to consider the following compared to a non-generative alternative:

- Will inclusion of generative AI increase the risk class of medical equipment?
- Could a product with generative AI provide misinformation and thus pose a risk to public health?

- Is generative AI appropriate for the intended use area?

Source: https://www.fda.gov/media/182871/download

Inadequate competence in dialogue with language models

Queries, also called prompting, using natural language makes it easy for users with different competence and backgrounds to use large language models. However, responses can vary based on the exact wording of the question asked, and it is therefore not irrelevant how questions are formulated [58].

Certain phrasings, imprecise queries, and insufficient description of context can give incorrect or inaccurate responses.

Prompt and system prompt

Prompt: A text-based instruction sent to a language model (like ChatGPT) to get a response. A prompt is the user's question or request, and the quality of a prompt directly affects the relevance and quality of the response.

System prompt: An overarching instruction that defines the language model's role, limitations, and behavior. A system prompt is usually not visible to the end user but controls how the model interprets and responds to all user prompts. A system prompt can, for example, define that the model should act as a health professional assistant, respond concisely, and always refer to professional guidelines.

Knowledge about dialogue with language models, along with practical training can, however, improve the quality of responses from language models. For example, a study has shown that structured prompts gave better results than unstructured ones [59] and that two-step reasoning (first asking for a summary, then asking for diagnosis) gave better diagnostic accuracy than simple prompting [60]. A study demonstrated how more advanced models (like GPT-4) handled complex prompts well, while weaker models (like GPT-3.5) performed worse with such prompts on clinical reasoning tasks [61]. Another study has shown that large language models had limited utility for clinicians as decision support tools, but the authors suggested greater potential if users receive good training in phrasing of good prompts [62].

Reduced professional knowledge and skills

Increasing use of AI and large language models in health and care services can lead to gradual dependence on the tools. Healthcare personnel and patients may come to rely on the technology, which in the long term can affect both the ability to make independent assessments and acquired skills.

Confirmation bias (algorithmic appreciation) means that users unconsciously weight information that confirms their existing beliefs [63]. Both the phrasing of a prompt and interpretation of the response given by a language model can be influenced by confirmation bias. This can be challenging when using AI tools in health and care services, as healthcare personnel may become more inclined to accept responses from a language model if it aligns with prior expectations [64]. Consequently, there is a risk that incorrect responses go undetected, especially if the model presents information with great confidence, both with and without source references. Training in critical use of AI tools will be crucial to reduce this risk.

Automation bias is excessive trust in AI tools that can impair professional integrity and quality. As AI systems improve and give correct answers more frequently, it will become harder for healthcare personnel to catch the occasional wrong answers. This can lead to users not questioning or considering other sources of information, because they assume the AI system is correct.

Deskilling. Several sources, including WHO and the National Health Service (NHS) in England, point to deskilling as a risk with extensive use of AI systems in health services [65][66]. Loss of skills, knowledge, or weakened decision-making ability can occur if skills are not in regular use, such as if tasks are left to machines, including language models. The result can be that healthcare personnel do not check or challenge a decision proposed by a model. They may also become unable to perform certain types of tasks or procedures in cases where the model is unavailable, for example during network failures or security breaches.

An international survey among healthcare personnel shows that a large fraction are concerned that use of generative AI can weaken critical thinking and lead to increased dependence on AI in clinical decisions [67] . Loss of skills will also apply to the next generation of healthcare personnel who will increasingly encounter AI tools as part of education, training, and changes in clinical practice [68]. Increased experience with using large language models will lay the foundation for new research that may eventually change our understanding of this risk.

Reduced privacy and anonymous information

There will be risk related to compliance with privacy requirements when large language models are in active use. If sensitive personal data is entered into a prompt to a language model that stores data and/or learns continuously, for example in a web application, there is a risk that sensitive data is used to train and update the model, and that it can go astray.

One way to secure language models that handle personally sensitive information can be to store and run the model in private cloud services or locally (on-prem solutions). Settings that ensure encryption of data (in use) and controlled (or lack of) data storage are other methods to ensure that input or output data is not exposed.

A common technique for complying with privacy requirements is to only share and process anonymous information. The risk to privacy is reduced if personal data is anonymized because the data subjects can no longer be recognized in the dataset. Anonymous information cannot be linked to an individual and is therefore not personal data regulated by the General Data Protection Regulation (GDPR).

Personal information is considered anonymized when it is handled or processed so that it can no longer be linked to an identified or identifiable natural person. In assessing whether the information is anonymous or not, one must see if it is possible to trace the information back to the individuals the information relates to.

Datasets for anonymization must be processed to avoid possibilities for re-identification (method for finding back to a person's identity from anonymized data) or reverse identification (method for re-identification, often using combination with other data sources). Assessing whether the information in a dataset is to be considered genuinely anonymous as a result of the anonymization process will depend on a comprehensive and risk-based assessment that rests with the data controller. Real anonymization can be difficult to achieve for certain datasets, for example for very comprehensive datasets with many variables.

With current access to large amounts of data and powerful analysis technology, it will be possible to a greater extent than before to re-identify individuals through information in the dataset [69]. This risk is important to assess when sharing information.

Complex regulations

Development and use of artificial intelligence, including large language models in health services, must always occur within the framework of applicable law. Users (deployers) as well as developers and providers of AI systems to be used in health and care services in Norway must comply with a range of laws and regulations, including both general and sector-specific regulations. For example, sector-specific regulations such as Norwegian health legislation and the EU regulation on medical devices apply. Furthermore, general regulations such as the EU's General Data Protection Regulation, copyright law, and equality and anti-discrimination law apply. Additionally, the AI Act, which has entered into force in the EU and should be implemented in Norwegian law as soon as possible [70].

The relationship between the various and sometimes overlapping regulations is complex. The complexity can lead to different regulatory understanding and interpretation, that processes take time, or that the scope for action in the regulations is not fully utilized.

Uncertainties about the legal scope for action within applicable law can lead to non-compliance. This may result in businesses, healthcare personnel, or citizens using AI systems with large language models in violation of regulatory requirements, or conversely, adopting overly restrictive interpretations that don't fully utilize the legal scope for action. For example, there may be uncertainty about whether an AI system should be classified as medical equipment or not, and if so, which risk class it belongs to [71].

Risk class according to medical device regulations is leading for which risk class the tool gets according to the AI Act, and thus decisive for which requirements apply to, among other things, quality assurance and documentation according to both laws. AI systems classified as medical devices often [72] also become classified as high-risk systems according to the AI Regulation [73]. Uncertainty about AI Act requirements therefore arise for AI tools where it is unclear whether they are covered by the definition of medical device, or which class of medical device applies. Examples of this can be chatbots used in therapy, or applications that use generative AI to create drafts for patient records [74].

Unintended illegal use

General purpose large language models can be used for a range of tasks, also by healthcare personnel and patients. Because the technology is easily accessible and simple to use, there is a risk that AI systems are used for purposes they were not intended or regulated for. Some examples are given below.

Illegal use of large language models in health and care services. If healthcare personnel are to use an AI tool for medical purposes, they must according to the handling regulations use CE-marked equipment [75]. Without quality-assured AI tools (such as CE-marked products), healthcare personnel may use large language models for purposes they were not developed for, making such use potentially illegal. Impatience, ignorance, or lack of clear guidelines can lead to healthcare personnel using language models for tasks that are not in line with the area(s) of intended use as defined by manufacturers or providers of AI tools.

Irresponsible use of AI tools by citizens. Citizens and patients have easy access to AI tools marketed as lifestyle tools, and AI models for general purposes that do not have specific medical purposes. Citizens may still use the tools for, for example, self-diagnosis or to get medical advice. This use can create risk of

incorrect or misleading information and/or that the information is misinterpreted. Such misinformation can affect health decisions and in the worst case threaten patient safety if they are not in contact with health services [76].

There is also risk related to illegal use of health information for training large language models. Section 29 of the Healthcare Personnel Act allows dispensation from confidentiality obligations for using sensitive information to train large language models for specified purposes [77]. If the model is later used for other tasks, which means using health information for a purpose other than what the dispensation decision covers, a legal basis for using information for the new purpose is required. Failure to ensure this can result in use of health information for new purposes without sufficient legal basis.

Risk of discrimination

Unrepresentative training data for large language models can introduce biases that risk discrimination in health and care services, a concern also highlighted by WHO in its report [78]. If the biases affect someone's right to benefits or access to health services, it can constitute illegal discrimination. AI systems that may pose this risk fall under the definition of high-risk AI systems in in AI Act [79]. Under the AI Act, both public and private users (deployers) of high-risk AI systems in such cases must perform a Fundamental Rights Impact Assessment (FRIA). This is an assessment of how an AI system can affect human rights, at both individual and group levels [80]. If use of an AI system reveals risk of discrimination against a special group of patients, one measure could be to offer this group an alternative without AI.

Environment and sustainability

Development, testing, and use of large language models often require high energy consumption and will leave a significant environmental footprint. Energy consumption is linked to pre-training and post-training of models [81], and to use [82]. The environmental impact also includes high consumption of water for cooling large data centers and extraction of rare minerals necessary for today's hardware used for model training [83][84].

The health and care sector is estimated to account for 4-5% of global greenhouse gas emissions, and both procurement and use of equipment and information technology contribute to emissions. The Norwegian Directorate of Health has published a roadmap containing concrete measures for how health and care services can become more environmentally friendly [85]. AI is highlighted as a technology that, used wisely, can facilitate a more sustainable health service, despite its considerable resource demands.

There is a requirement to weight climate and environment at 30% in new public procurements where relevant [86]. In practice, however, it can be difficult to get an overview of the climate footprint when procuring AI solutions. Suppliers often provide little information about energy consumption and carbon emissions, and there is no common standard for how such calculations should be conducted. Under Article53 of the AI Act, providers of large language models to be used in the EU are required to include computational resource and energy consumption information in their technical documentation. The Transparency chapter of the voluntary Code of Practice provides a guides and template to this requirement [87].

As an alternative to large language models, small models are also being developed that require significantly less energy and resources both for training and in use [88]. Using techniques such as model distillation [89] and targeted fine-tuning, these models perform increasingly better and can be a good alternative for solving various tasks [90].

Scientific evidence and responsibility

Implementation of AI tools with consequences for patients' medical treatment requires caution. The consequences of using the tools are comparable to new medications or other medical equipment. It will be necessary to develop new tests and areas for evaluation of large language models to be used in health, as well as studies and documentation [91][92]. For example, WHO recommends conducting randomized clinical trials to prove that an AI system to be used in clinical practice has better performance than alternatives, not only perform testing in a laboratory or controlled environments [93]. Without a good scientific foundation, proper training, and correct implementation, the burden of responsibility will be unclear if errors occur and healthcare personnel have relied on responses from AI tools. Such situations can have major ethical and legal consequences. Note that business leaders are responsible for facilitating correct implementation, training, and organizational measures.

The European Commission recognized challenges related to liability issues when using artificial intelligence and proposed a separate AI Liability Directive in 2022 [94]. The goal of the directive was to provide increased legal certainty for both patients and healthcare personnel if damage should occur as a result of using AI tools. In February 2025, the proposal was withdrawn due to lack of agreement among member countries, but the issue is still equally relevant.

[56] https://www.who.int/publications/i/item/9789240084759

[57] https://www.fda.gov/media/182871/download

[58] https://arxiv.org/abs/2311.16452

[59] https://www.medrxiv.org/content/10.1101/2024.09.01.24312894v1.full

[60] https://www.medrxiv.org/content/10.1101/2024.09.01.24312894v1.full

[61] https://arxiv.org/abs/2308.06834

[62] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395#zoi241182r32

[63] https://journals.sagepub.com/doi/10.1037/1089-2680.2.2.175

[64] https://ai.nejm.org/doi/full/10.1056/AIe2400961

[65] https://www.who.int/publications/i/item/9789240084759 p.12

[66] https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/developing-he

[67] https://assets.ctfassets.net/o78em1y1w4i4/kWTSca6VXZ54DBhAIYxJU/386d36dc0c03c4fa8de0365bbb20

[68] https://pubmed.ncbi.nlm.nih.gov/38662419/

[69] https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/informasjonssikkerhet-internkontroll

[70] https://www.regjeringen.no/en/dokumenter/the-digital-norway-of-the-future/id3054645/ p.71

[71] https://www.scup.com/doi/10.18261/olr.10.1.1

[72] Medical devices of class iia or higher according to MDR

[73] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689#cpt_III.sct_1 Aticle 6

[74] https://www.scup.com/doi/10.18261/olr.10.1.1

[75] https://lovdata.no/dokument/SF/forskrift/2013-11-29-1373

[76] https://www.who.int/publications/i/item/9789240084759 p.15

[77]
https://www.helsedirektoratet.no/rundskriv/regelverket-for-utvikling-av-kunstig-intelligens/regler-for-de-ulike-

[78] https://www.who.int/publications/i/item/9789240084759 p.xi, Table 2, p.21

[79] https://artificialintelligenceact.eu/annex/3/ 5.(a)

[80] https://artificialintelligenceact.eu/article/27/

[81] https://hbr.org/2023/07/how-to-make-generative-ai-greener Training GPT-4 is estimated to have generated around 300 tons of CO emissions, compared to an estimated 5 tons annually for a human.

[82] According to the ecological impact calculator available on Hugging Face it is estimated that one query to GPT-4o with a relatively short response (400 tokens) consumes 35.1 Wh of electricity, approximately seven times more than charging a mobile phone.
https://huggingface.co/spaces/genai-impact/ecologits-calculator visited February 2025

[83] https://www.nature.com/articles/s41598-024-76682-6

[84] https://arxiv.org/pdf/2304.03271

[85]
https://www.helsedirektoratet.no/rapporter/veikart-mot-en-baerekraftig-lavutslipps-og-klimatilpasset-helse-o

[86]
https://anskaffelser.no/verktoy/veiledere/veileder-til-regler-om-klima-og-miljohensyn-i-offentlige-anskaffelser

[87] https://artificialintelligenceact.eu/annex/11/ and
https://code-of-practice.ai/?section=transparency#model-documentation-form

[88] https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai

[89] https://www.ibm.com/think/topics/knowledge-distillation

[90] https://arxiv.org/abs/2305.02301

[91] https://ai.nejm.org/doi/full/10.1056/AIe2401235

[92] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf p.28

[93] https://www.who.int/publications/i/item/9789240029200 p.107

[94] https://www.ai-liability-directive.com/

Summary

The use of generative AI and large language models is currently an immature field. Although the possibilities are many, there is also great uncertainty about risks related to how this type of AI can be used in a trustworthy way in health and care services. We simply lack insight into the side effects.

However, the field is developing rapidly, with among other things mechanisms to improve, ground in knowledge, and gain better control over the quality of AI systems with generative language models.

One central measure to reduce risks and increase quality is to develop and make available language models, i.e., foundation models, that are adapted to conditions in the Norwegian health and care sector, and that industry and the public sector can build on for their use. This is in line with the national digitalization strategy which states that the Government will work to further develop a national infrastructure for AI that will provide access to foundation models based on Norwegian and Sami languages and social conditions [95].

The next chapters cover how large language models can be adapted to Norwegian health and care services (Chapter 4) and propose measures that together will facilitate the use of generative AI models being based on foundation models adapted to conditions in the Norwegian health and care services (Chapter 5).

[95] https://www.nb.no/pressemeldinger/regjeringen-vil-satse-stort-pa-digitalisering-og-kunstig-intelligens-ved-na In the State Budget for 2025, 20 million kroner is allocated to increased investment in the National Library's work with artificial intelligence and training of generative language models.

Adaptation to the Norwegian health and care services

For a generative language model to be used safely and effectively in Norwegian health and care services, it should be well-adapted to Norwegian conditions. In this chapter, we will look at *what* adaptation of language models to Norwegian conditions entails, *how* language models can be adapted to Norwegian conditions, *what is needed* to adapt language models to Norwegian conditions, and *who* can adapt the language models.

This report does not address general adaptation to Norwegian conditions in detail, such as language in Norway or general public administration and knowledge about social conditions and values but focuses primarily on adaptation to the Norwegian health and care services. In the general public domain, we will seek cross sectoral collaborations.

Adapting to Norwegian conditions does not exclude that the language model should also be adapted to international conditions. Health and care services are provided in an international context, and health professional knowledge and innovations occur internationally. A language model should not be adapted to Norwegian conditions alone.

Large language models can be components in AI systems. AI systems can build on one or more different language models and usually have an intended area of use.

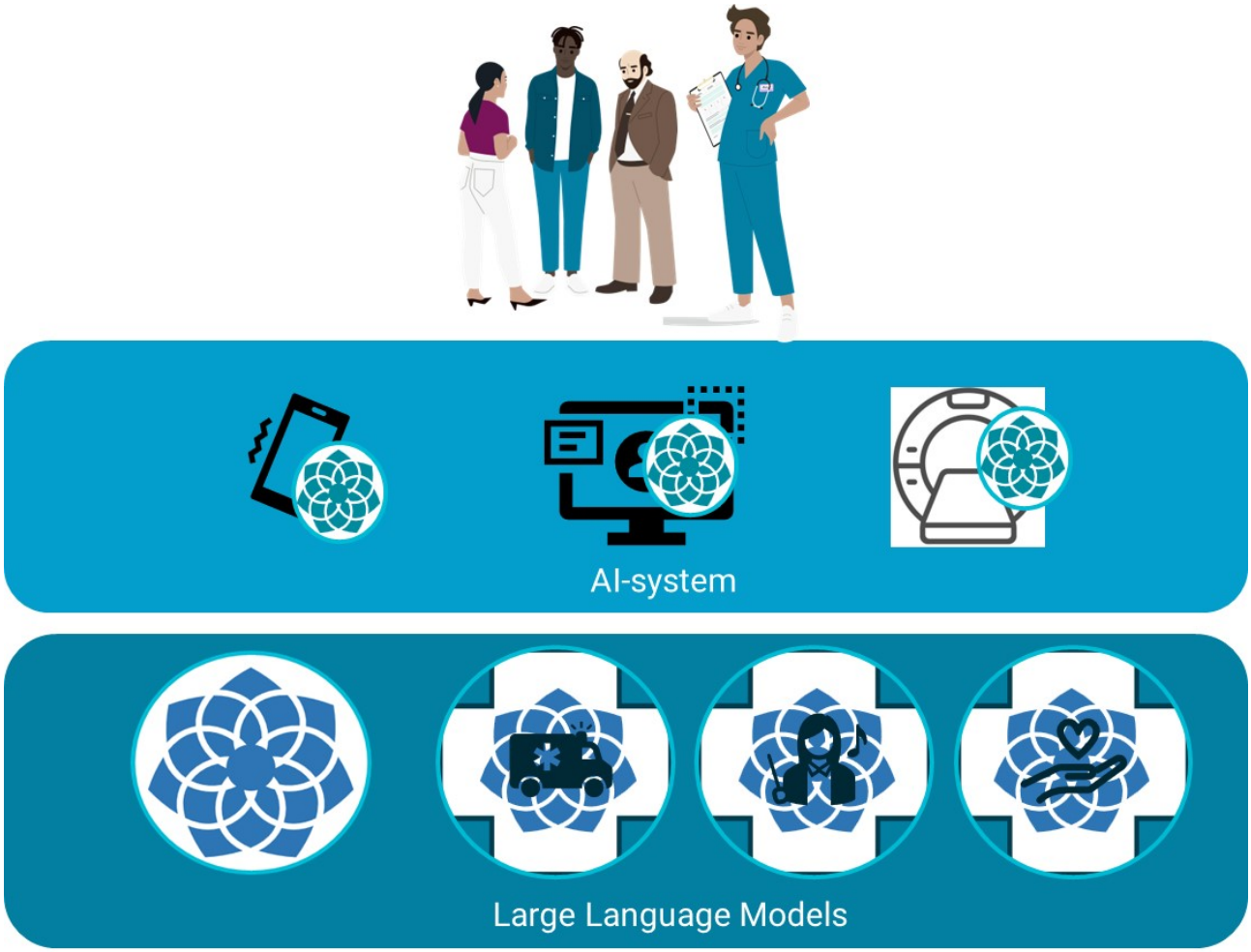

Figure 3: Large language models can be components in AI systems. AI systems can be built on one or more different language models and usually have an intended use area.

What is adaptation to Norwegian conditions?

Norwegian conditions is a broad and vague concept that can encompass very much. Some will apply generally for use in Norway, but much will apply specifically to health and care services in Norway. We will look more closely at the latter.

Based on acquired knowledge, adaptation to Norwegian conditions can be divided into five areas: languages in Norway, general knowledge in Norway, public administration in Norway, legislation in Norway, values and ethics in Norway. Similarly, adaptation to Norwegian health and care services can be divided into five areas: health professional language in Norway, health professional knowledge and practice in Norway, health administrative knowledge and practice in Norway, Norwegian health legislation, and values and ethics that apply in health and care services [96].

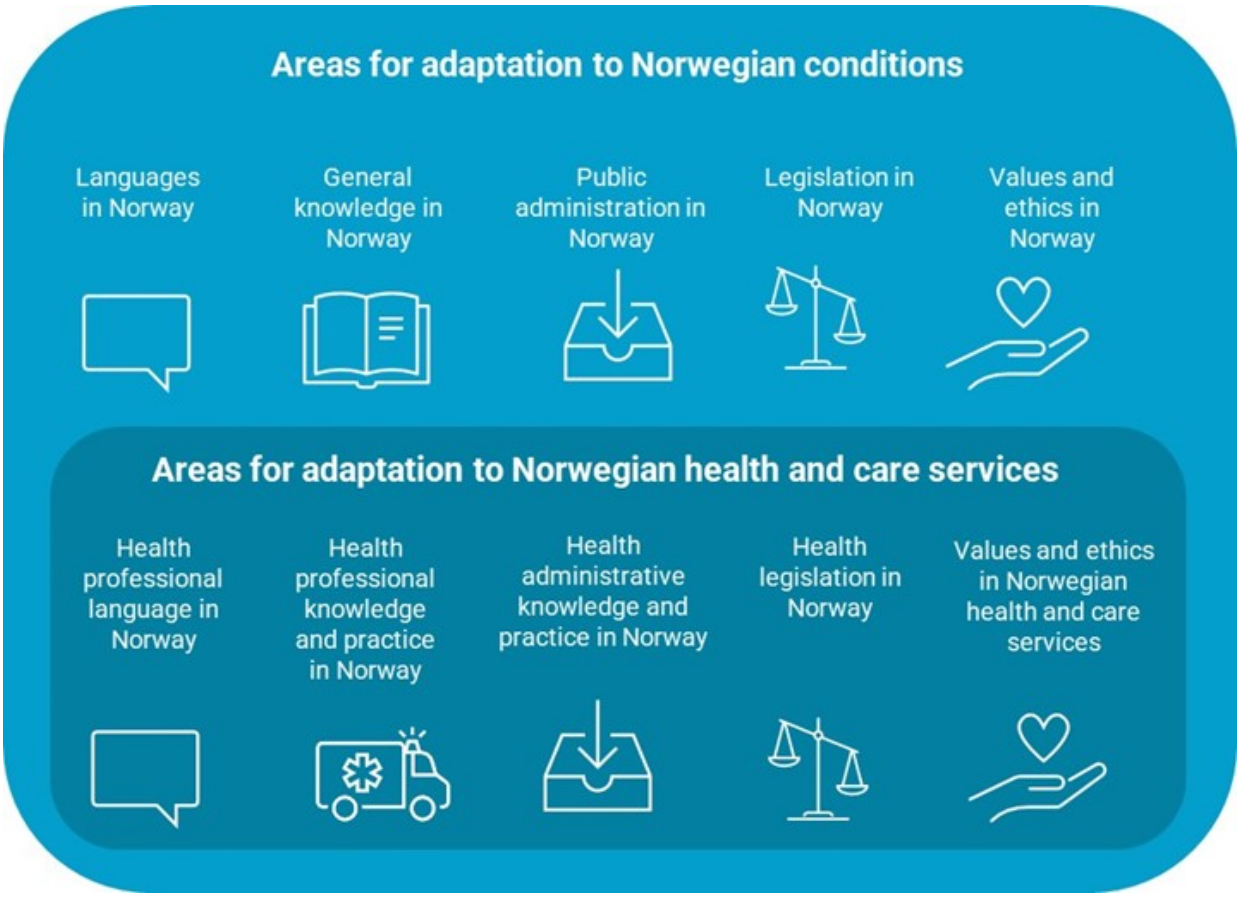

Figure 4: Language models can be adapted to different areas: to Norwegian conditions generally and for health and care services specifically. Areas for adaptation concern language, professional knowledge, administrative knowledge, legislation, and values and ethics.

Languages in Norway

Bokmål and Nynorsk

Language models adapted to Norwegian conditions must be able to perform tasks in Norwegian. The Language Act provides several guidelines for public language use in Norway. According to Section 4 of the

Language Act, Norwegian is the national main language in Norway, and Bokmål and Nynorsk are equal written languages in public bodies [97]. In addition, public bodies shall comply with the official orthographic rules for Bokmål and Nynorsk. Language models for use in public health and care services must therefore master Bokmål and Nynorsk within the established orthography. Language models should also be able to handle variation within Bokmål and Nynorsk.

The Language Council states in a linguistic survey from October 2024 that the language in ChatGPT did not maintain sufficiently high quality to be used by public administration [98]. Evidence suggests that language models are still not well enough adapted to Norwegian language.

Sami Languages

Sami languages, i.e., Lule Sami, South Sami, and North Sami, have status as indigenous languages. According to Section 5 of the Language Act, Sami languages and Norwegian language are equivalent [99].

The purpose of the Sami Act is to create conditions for the Sami ethnic group in Norway to secure and develop its language, culture, and community. It sets requirements for when public bodies must use Sami languages. At the same time, the law gives individuals linguistic rights in meetings with public bodies. Section 3-5 of the Sami Act gives persons extended right to use Sami in health and care institutions in the administrative area for Sami language [100][101][102].

Work should be done to develop and make available language models that function in Sami taking into account both the linguistic and cultural perspective.

Other Languages in Norway

Norway is a multilingual country, and we find a range of other languages in use in addition to Norwegian and Sami languages: the national minority languages, Scandinavian, Norwegian sign language, and newer minority languages.

North Sami, South Sami, Lule Sami, Kven, Romani, and Romanes are all defined as minority languages in Norway and thus protected by the minority language charter [103]. Speakers of these minority languages do not have their own concrete rights related to meetings with health and care services. As with Sami language users, efforts should be made to ensure that AI tools consider the cultural perspective in meetings between users from minority groups and society at large, which will be in line with the Patient and User Rights Act. In the health and care sector, this will largely apply to the newer minority languages.

Scandinavian languages can be used in patient records. According to Section 10 of the Patient Record Regulation, Danish and Swedish can be used to the extent that it is responsible [104].

Multilingual Language Models

Norwegian health and care services use an intricate mix of Norwegian and international medical terminology. This requires advanced embedding spaces that can handle multiple languages simultaneously.

An embedding space is a multidimensional mathematical space where words, sentences, images, or other data are represented as vectors. Similar concepts are placed near each other based on semantic or contextual similarity.

Modern language models implement this through shared vector spaces for medical terminology, while maintaining separate representations for Bokmål and Nynorsk with a common medical vocabulary as

foundation. Extensive testing has shown that this architecture, especially with dedicated medical embedding levels, results in a significant increase in precision in cross-linguistic medical communication.

Source: https://link.springer.com/article/10.1007/s10462-025-11162-5

Section 6 of the Interpreting Act gives public bodies a duty to "use an interpreter if this is necessary to uphold legal safeguards or provide proper assistance and services. In the assessment of whether it is necessary to use an interpreter, weight shall be given to factors such as whether the parties to the conversation can communicate adequately without an interpreter, as well as the gravity and nature of the matter." If language models are to be used in such cases, for example for automatic interpretation, the Interpreter Act will probably apply. Language models must then be adapted to relevant languages.

Norwegian health professional language

If language models are to be used in health professional contexts, they should in addition to Norwegian master the Norwegian health professional language, including professional terminology. Professional terminology includes specialized professional terms, abbreviations, and jargon.

The professional language is characterized by many synonyms, where Norwegian, Greek, Latin, and hybrid (Latin terms adapted to Norwegian spelling) are often found in combination. In addition, professional terminology varies greatly between different professional groups and specialties.

Language models should therefore be able to both receive instructions and produce text with established and correct Norwegian terminology for the relevant area of use. It would, for example, be unfortunate if a language model automatically translates and creates completely new direct translations where there are well-established professional terms that have standing in Norwegian professional language. This has so far been a challenge when using AI tools for translation.

Dialects

If language models are to be used to process speech, for example speech-to-text, the language models should also master the dialects in use in Norway as well as other oral speech variations such as Scandinavian languages. The use of dialects has been a significant challenge with previous language technology. However, newer generative language models like NB-Whisper have shown promising results. This becomes particularly important for several types of tools, for example speech-to-summary (ambient scribe solutions) and oral language use in chatbots (conversational AI).

NB-Whisper

The National Library has adapted the language model Whisper from the company OpenAI to Norwegian. It is a language model that can convert speech to text, for example conversations and monologues.

NB-Whisper has been trained on large amounts of language resources from the National Library, including Norwegian speech corpus. This makes the language model understand Norwegian speech and Norwegian dialects far better than previous technology.

Like other generative models, NB-Whisper can also experience hallucinations, which is particularly important to be aware of in health and care services.

NB-Whisper is freely available for use, also in the health and care sector.

NB-Whisper is a good example of how an international language model can be adapted to Norwegian conditions. Systematic work with building good language resources at the foundation has facilitated that the public or industry can develop solutions for health and care services. Several technology suppliers now offer speech-to-text technology based on NB-Whisper.

NB-Whisper is made available by the National Library, which manages and further develops the model.

Source:
https://www.nb.no/pressemeldinger/nasjonalbiblioteket-deler-kunstig-intelligens-som-skjoner-norske-dialekt

Plain language

The Language Act sets requirements for plain language, i.e., clear and correct language adapted to the target group. In addition, Section 3-5 of the Patient and User Rights Act is relevant: When information is to be given about a patient or user's health condition and health care, the information shall be adapted to the recipient's individual prerequisites, including culture and language background. Language models should therefore be able to differentiate language use and choice of words depending on who the users/readers are, whether they are citizens with different levels of health literacy or professionals.

Difference between languages is not always about different words and sentence structures. Language is also a cultural expression. Norms for good language use, such as politeness, style, and tone, vary between cultures and languages. An adapted language model should therefore also be able to express itself in a way that is in line with norms for good language use in Norway.

Health professional knowledge and practice in Norway

Health professional knowledge can be international and independent of countries and cultures. At the same time, much health professional knowledge is conditioned by specific cultural areas. For example, knowledge about diet and nutrition and associated advice can vary from country to country, and one often speaks of Nordic nutrition recommendations [105]. A language model adapted to Norwegian conditions must be able to build on Norwegian (or Nordic) knowledge and our health professional understanding of the world, as the nutrition example illustrates.

Language models used in health and care services should also work in line with best practice in the service. There are a range of national professional guidelines and patient pathways that represent standardized clinical practice in Norway, for example antibiotic guides, cancer care packages, and guidance plans for nursing practice. A prerequisite for safe use of language models in Norway is precisely that it is trained on data containing such national guidelines.

No systematic analysis has been done of the extent to which available language models are based on such knowledge.

Without specific training or guidance, language models can fall back on general or international knowledge rather than following country-specific guidelines and practices.

Practices and guidelines can be clarified and defined through standards, and they can help language models generate content that is in line with Norwegian conditions. An international survey may suggest that integration of standards in language models can improve their compliance with such standards [106].

Health administrative knowledge and practice in Norway

The Norwegian health and care sector is structured differently than in other countries. This applies to, among other things, the distribution between private and public health and care services and the geographical and administrative orientation. Furthermore, it also applies to arrangements such as general practitioners and the distinction between primary and specialist health services in general, as well as the central health administration [107].

A language model adapted to Norwegian health administrative knowledge and practice should be able to handle knowledge about this. It is currently uncertain to what extent available language models are capable of this [108].

It is important to recognize that health administrative knowledge and practice evolves differently than health professional knowledge, changing primarily through political decisions and legal modifications.

Health legislation in Norway

Rights and obligations of patients, relatives, healthcare personnel, and businesses vary from country to country. A language model adapted to Norwegian conditions also entails adaptation to Norwegian legislation.

A range of laws and regulations regulate the health and care sector, from general laws like the Health and Care Services Act to more specific laws like the Abortion Act. It is absolutely crucial that a language model in use in Norway does not provide information that violates either Norwegian law or Norwegian health law. No systematic analysis has been done of whether available language models are based on such knowledge.

Values and ethics in Norwegian health and care services

No language model is completely neutral when it comes to cultural preferences, ideology, and values. Research indicates that large language models reflect cultural values from training data. What utterances are perceived as offensive or aggressive will vary from culture to culture and will be based on the datasets used for training [109]. New large language models from other cultural areas than Western culture have also led to greater attention related to censorship and sensitivity and sparked a public debate. Commonly known, but politically charged, political events in China appear to have been censored in DeepSeek's language model [110]. Views on and knowledge about gender will, for example, vary between different countries and cultures.

The same will also apply to the health and care sector. In White Paper No. 26 (1999–2000) On Values for the Norwegian Health Service, it is stated that inviolable human dignity is the fundamental value for the service. Furthermore, equality, justice, equal access to services of good quality, professional responsibility, human dignity, and solidarity with the most vulnerable are central values.

This is also stated in Health Services in Norway that "[i]t is a fundamental principle that everyone in Norway shall have the same right to health services, regardless of age, gender, place of residence, and economy." [111].

In several white papers in recent years, emphasis has been placed on user participation as a fundamental prerequisite for good health and good care [112]. Patients, users, and relatives shall be seen and heard. User participation is a fundamental prerequisite in the patient's health service, where no decisions shall be made about you without you. It shall be based on the person's holistic needs, including physical, mental, social, spiritual, and existential needs.

Such values are not necessarily universal, and it must be ensured that language models in health and care services are in line with this value foundation.

[96] https://tidsskriftet.no/2024/11/kronikk/kritisk-veiskille-integrering-av-kunstig-intelligens

[97] https://lovdata.no/dokument/NLE/lov/2021-05-21-42

[98] https://sprakradet.no/aktuelt/ki-sprakets-fallgruver/

[99] https://lovdata.no/dokument/NLE/lov/2021-05-21-42

[100] https://sametinget.no/sprak/forvaltningsomradet-for-samiske-sprak/

[101] The new regulation implementing the language rules in the Sami Act will expand the range of nationwide organizations that Sami language users have the right to receive written responses from.

[102] https://lovdata.no/dokument/NLE/lov/1987-06-12-56/KAPITTEL_3#KAPITTEL_3

[103] https://www.regjeringen.no/no/tema/urfolk-og-minoriteter/nasjonale-minoriteter/midtspalte/minoritetssprakpa

[104] https://lovdata.no/dokument/SF/forskrift/2019-03-01-168?q=pasientjournalforskriften

[105] https://pub.norden.org/nord2023-003/recommendations.html

[106] https://arxiv.org/abs/2402.12593

[107] i.e https://www.akademika.no/medisin-helse-og-psykologi/medisin-og-medisinske-disipliner/helsetjenesten-i-nc

[108] ChatGPT 4o was queried in January 2025 about hospitals in Møre and Romsdal. When asked in English, it listed all four hospitals, but when asked in Norwegian, it only listed three hospitals, Kristiansund Hospital was omitted.

[109] https://arxiv.org/html/2402.10946v1 and https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548

[110] https://www.kode24.no/artikkel/4-grunner-til-a-vaere-skeptisk-til-deepseek/82594605 , https://www.khrono.no/fastlaste-verdier-for-alltid/939158

[111] Nylenna, Magne 2024 p.33

[112] i.e https://www.regjeringen.no/no/dokumenter/meld.-st.-38-20202021/id2862026/

How to adapt to Norwegian conditions?

How language models are developed and adapted is covered, among other things, in the report "Large Language Models in Health and Care Services." [113]

Methods for adapting language models are not mutually exclusive. Pre-training involves training a model from scratch so that it can be adapted to Norwegian conditions from the very beginning. Post-training involves building on a pre-trained general model with data sources relevant to the health sector. An AI system that uses language models can also be improved and adapted to concrete use, for example through knowledge anchoring or through agentic AI. This chapter describes various ways to pre-train, post-train, knowledge-anchor, evaluate, and test large language models.

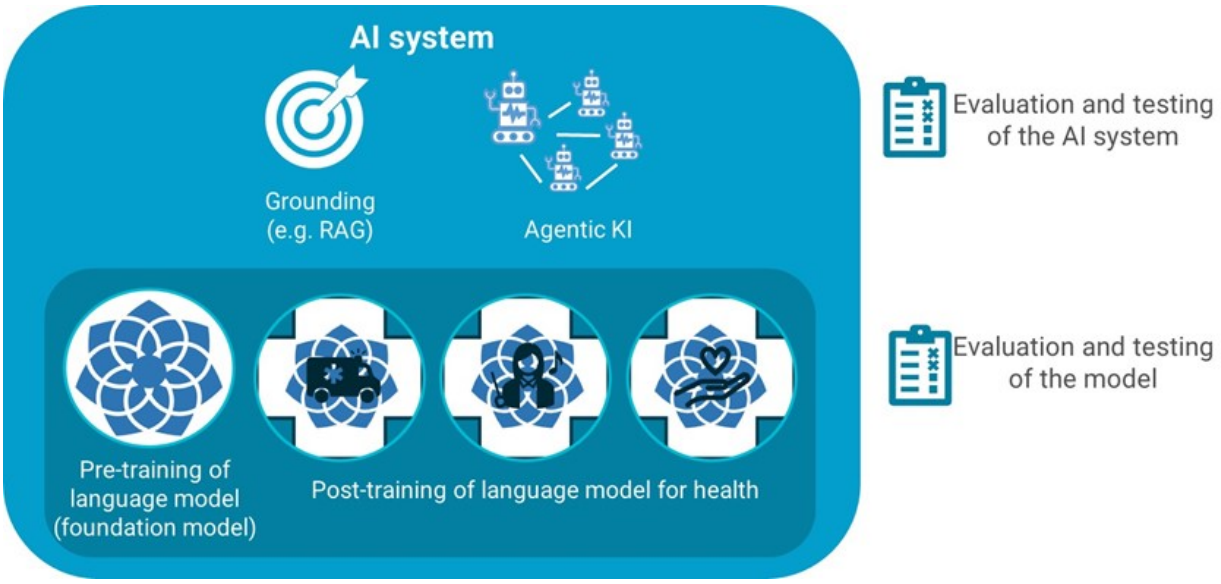

Figure 5: Adaptation to Norwegian conditions can be achieved through pre-training of a large language model and/or post-training of pre-trained language models. An AI system using language models can be improved and adapted to specific use through for example knowledge grounding or through so-called agentic AI. Both the AI model and the AI system must be evaluated and tested.

Pre-training of language models

Pre-training of large language models gives full control over the model and insight into the model's learning process. However, the approach is very costly, as it will require extensive amounts of relevant data and significant computational resources. GPT and Gemini are examples of large, market-leading pre-trained language models that have been developed from scratch. These are also called foundation models and are often delivered by large American technology companies. However, challenges are pointed out related to, among other things, explainability, accountability, sovereignty, and linguistic quality in these models [114]. As a response, several European countries have initiated developing their own pre-trained language models.

The large international foundation models have proven to be powerful and applicable in many areas and already contain some Norwegian language and are used in Norway. In Norway, there are several initiatives

to develop Norwegian foundation models, both through BERT-based models like NorBERT and larger models like NorGPT from NorwAI and NORA.LLM models from the NORA consortium, for example NorMistral.

Many large international foundation models are pre-trained on significant amounts of general health professional knowledge. There are also international language models that are pre-trained from scratch specifically for health. An example is MedFound, which is trained on patient records and medical texts [115]. This model is much smaller than the larger international foundation models mentioned above. Such smaller models can be more energy efficient, but the areas of use is more limited since they are trained on less data.

The technology is rapidly evolving, and new models such as DeepSeek's language models R1 and Open R1 represent a new and more energy-efficient direction for how information can be retrieved and processed. This makes previous approaches, particularly BERT-based models, largely outdated for generative tasks, although they may still be useful for specific applications.

When choosing which pre-trained model to build on, it is important to consider whether it is open or closed. For open models, either the source code can be available and/or the training data can be known and documented. Conversely, access to source code and/or training foundation will be limited or not available for closed models. This makes it difficult, if not impossible, to explain and quality assure models derived from them.

Post-training of pre-trained language models (foundation models)

Although many foundation models already master some Norwegian language and are trained on some international health professional knowledge, they should be post-trained if they are to become better adapted to Norwegian conditions. Post-training involves building on a pre-trained general model with relevant data and can be based on one or more foundation models.

The National Library's NB-Whisper and Bineric's NorskGPT are examples of Norwegian, general (non-health professional) post-training.

Fine-tuning with Norwegian health professional language resources

Foundation models can be fine-tuned using domain-specific health professional resources as training data. A fine-tuned language model will then be much better at handling Norwegian professional language and terminology as well as health professional and health administrative knowledge and practice.

A foundation model can be fine-tuned to:

- a relatively large and general health professional language model. This will facilitate a common health professional language model for Norwegian health and care services that can be used and further developed by the entire sector, including public and private actors.
- several smaller and specialized language models, oriented toward different purposes (subject, specialty, or task). These are trained on texts belonging to a given task or specialty.

Such fine-tuning can be achieved in several ways. One possible approach is to train the foundation model further on Norwegian language, health professional texts so that the algorithms in the model are changed.

This requires that the foundation model is open (see above). This form of fine-tuning requires significant computing power [116]. Another approach involves fine-tuning by adding layers outside the pre-trained model, which preserves the original model weights and therefore requires less computing power [117].

Regardless of approach, it is important to be aware that a language model will never be fully learned. Professional language, knowledge, and practice in health and care services are not constant but are updated in line with professional development in the field. It is therefore important to account for continuous learning when fine-tuning a language model. Methods for fine-tuning are an active area of research and can be facilitated by federated learning [118]. Another approach is to version and periodically release models with updated knowledge. This will require active management.

Continuous updating of language models

Large language models can be updated continuously by establishing a systematic follow-up and update mechanism so that, for example, new clinical guidelines, updated data, and changes in professional terminology can be quickly incorporated into the model. This can, for example, involve regular updates through continuous learning or periodic fine-tuning.

A feedback loop with clinical expertise can also be established, for example by including a mechanism for continuous feedback from health personnel using the model. The feedback is used to identify any errors or outdated recommendations so that the model can be revised and improved in line with current professional knowledge and practice.

Instruction tuning in line with health professional and administrative tasks

To handle health professional tasks, a language model must be specifically trained for them. Such tasks can be translations, interpretation, conveying information to patients, or creating drafts for patient records. This can be done by fine-tuning the language model using datasets containing examples of the tasks to be solved. The process can be iterative and involve humans providing feedback along the way.

The need for instruction tuning applies to both foundation models and fine-tuned models. Even if models are instruction-tuned in English beforehand, there may still be a need for specific Norwegian instruction tuning. This is particularly relevant for domains where precise understanding of Norwegian instructions is critical, or where cultural and subject-specific aspects are important.

Fine-tuning in line with Norwegian legislation, values, and ethics

A language model can be fine-tuned to function in line with Norwegian legislation, values, and ethical principles. Such adaptation can be done in several ways:

- through systematic training on examples that demonstrate desired ethical behavior
- using reinforcement learning with human feedback (RLHF) where humans evaluate the model's responses against defined ethical guidelines
- by building ethical principles into training data and evaluation criteria

The purpose is that the model's behavior and generated text should reflect and be consistent with defined ethical principles. This should be seen in connection with a national approach to common values and ethics. For health and care services, values such as confidentiality, self-determination, and non-harm are central. At the same time, one must account for the fact that values and ethical principles are not constant but evolve over time.

When using foreign models, it is important to investigate what values and principles are already built into the model and assess how these interact with or possibly deviate from Norwegian values and health service needs.

To comply with legislation, the adaptation process can include integrated security and privacy strategies that explicitly include measures to protect sensitive health data.

Knowledge grounding and agentic AI

Knowledge grounding is a technique for improving the quality and relevance of responses from a language model without changing the language model itself. Quality-assured and curated data sources such as internal or external knowledge bases can constitute a knowledge foundation.

RAG (Retrieval Augmented Generation) is one method for knowledge grounding that has proven promising as an effective alternative to health professional fine-tuning [119]. A prerequisite is that the knowledge base is relevant to Norwegian conditions. At the same time, the language model used must be able to handle languages in Norway.

It is still too early to determine whether the combination of RAG and large language models works sufficiently well for the relevant areas of use in the Norwegian health and care sector. There is a need for more experience and knowledge in the field for RAG to be recommended for large-scale use in health and care services.

When a RAG solution is used, knowledge is retrieved from predefined knowledge bases. The language model's function is primarily to interpret and phrase the question/instruction to find correct information in the knowledge base and then generate the response after information has been retrieved. The approach makes direct adaptation to the purpose of use possible, requires little computational and data resources, and can reduce errors and fabricated responses. The Norwegian Directorate of Health is testing RAG for the information service HelseSvar.

Example: HelseSvar

The Norwegian Directorate of Health has tested RAG technology to generate suggestions for answers to health questions from young people related to various areas, such as contraception, tobacco, and mental health. The goal is to make the work of healthcare personnel more efficient when they respond to young people who make contact through ung.no.

Several different language models have been tested, together with a RAG solution that uses a knowledge bank with previous questions and answers, or articles published online.

In the final report, it was concluded that "RAG and enterprise data provide professionally strong results. But it is necessary to improve the linguistic adaptation, especially for Nynorsk and simpler linguistic expressions, to make the AI assistants more accessible and user-friendly for young people." This requires processing texts and formulations adapted to the target group in the enterprise data that forms the basis for the RAG solution.

Furthermore, it concludes: "In the short term, we recommend that the AI solution HelseSvar (which uses RAG) is primarily used as a support tool for health responders, rather than as a self-help tool for citizens. The RAG solution delivers correct answers in 80–90% of cases, which illustrates a high degree of precision and reliability. Although the solution responds very well to health-related questions, it happens that it

answers incorrectly due to misunderstandings of the context of the question, or unclear enterprise data. Human quality assurance is therefore necessary for health related questions."

After the report was published, the project continues to work with methods that improve response quality. This includes, for example, techniques for how responses are built up, especially through the use of agents that interpret the question, develop a plan for obtaining relevant information, and compose the response in a structured and controlled manner.

Source: "Report – HelseSvar. Concept study for an AI assistant for citizen-directed information." Internal report in the Norwegian Directorate of Health.

Agentic AI involves breaking down complex questions into smaller tasks and combining multiple tools that can answer tasks. The agent interprets the user's question and develops a structured workflow with specific steps to answer the question (see AI fact sheet Intelligence enhancement and control) [120].

Evaluation and testing of language models for Norwegian conditions

We lack national principles for testing, evaluation, and quality assurance of language models for Norwegian conditions, which is a challenge that applies to the entire public administration and not just the health and care sector.

Evaluation and testing can be done of both foundation models and fine-tuned models, among other things, to be able to compare and possibly choose the language model that is best suited for further training and adapted to the purpose of use. Evaluation and testing of an AI system is conducted to assess how well it performs in relation to the intended use.

To be able to assess a language model's performance requires a structured approach to evaluation of the various aspects described in section 4.1. A central aspect is language, where the model can be tested in how it masters Norwegian medical professional terminology, or how it masters Bokmål, Nynorsk, and in some contexts also Sami in health professional context. The precision level in health professional communication can also be assessed.

Contextual understanding is equally important, where tests can show how the model performs in accordance with Norwegian health service organization and structure. This can apply to Norwegian treatment guidelines and procedures, as well as local administrative routines and documentation requirements.

To ensure thorough evaluation, a stepwise approach to testing is possible. This can start with basic evaluation of language and security, proceed to domain-specific testing for the health sector, continue with use-area-specific testing, and context-specific testing in the implementation environment, then preferably as part of an AI system. If the language model is part of an AI system that falls under the Medical Device Regulation, specific requirements for testing and validation will be imposed according to this law [121].

[113] https://www.helsedirektoratet.no/rapporter/store-sprakmodellar-i-helse-og-omsorgstenesta

[114] https://www.nora.ai/news/2024/we-need-open-norwegian-language-models.html and https://teknologiradet.no/en/publication/generative-artificial-intelligence-in-norway/

[115] https://pubmed.ncbi.nlm.nih.gov/39779927/

[116] Full parameter fine-tuning https://arxiv.org/abs/2306.09782

[117] Parameter efficient fine tune-tuning (PEFT) https://arxiv.org/abs/2410.21228

[118] https://medium.com/@bhat_aparna1/federated-learning-with-large-language-models-balancing-ai-innovatio

[119] RAG (Retrieval-Augmented Generation) is an AI method that combines information retrieval with text generation to improve the quality and relevance of responses.

[120] AI fact sheets will be published here https://www.helsedirektoratet.no/digitalisering-og-e-helse/kunstig-intelligens/ki-faktaark

[121] https://lovdata.no/dokument/NL/lov/1995-01-12-6 and https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng

What is needed to adapt to Norwegian conditions?

Through work on the report, we have received input on what should be in place to adapt large language models and their use to Norwegian conditions in health and care services. The input includes access to data, computing power, quality frameworks, competence, and a governance regime. This chapter goes through these areas, which also form the basis for recommendations for measures in the Joint AI Plan (see Chapter 5).

Figure 6: Central areas that should be in place to adapt large language models to Norwegian conditions include data, computational power, quality frameworks, competence, and governance.

Sufficient data and data quality

To adapt language models to the Norwegian health and care sector, there is a need for significant amounts of high-quality data for both training and testing.

The digitalization strategy highlights the importance of better and easier access to data for the health and care sector: " The health and care services sector has a great deal of information that can be useful for developing AI, such as registry data, medical imagery and patient records. It must become easier for relevant actors to access health data to use it with AI. Improved and easier access to health data is important for the further development of our common health service, and for research and business development, but national security interests must be protected." [122]

A key factor for using data for adaptation to Norwegian conditions is data quality. We have described risks related to poor data quality in large language models in section 3.1.1. Data quality refers to how reliable, accurate, and applicable data is for a given purpose. High data quality is crucial for good analyses, decisions, and models.

Key factors for good data quality are:

- accuracy: the data correctly reflects reality
- representativeness: the data represents the entire population in Norway, including minorities and the Sami indigenous population
- completeness: no important data is missing
- consistency: the data is coherent across sources and systems
- reliability: the data comes from a credible source and is verifiable
- timeliness: the data is updated and relevant
- uniqueness: no duplicated or contradictory entries
- relevance: the data is useful for the purpose it is used for
- balance: the various health professional areas must be covered

Few data sources will be perfect, but various frameworks for describing datasets can be used to increase transparency around the data used to develop language models for use in the health and care sector [123]. Both Model Documentation Form for large language models related to the AI Act [124] and model cards from HuggingFace [125] have dedicated fields for describing data.

Collection and processing of data requires substantial effort. This also applies to clarifying whether there are legal, technical, or other limitations.

One can distinguish between labeled data and unlabeled data. Labeled data often has higher quality since it can contain semantic or contextual information that gives added value.

The question of availability of necessary data has been significant for several informants in this report.

Data categories

Data for language models can overall be divided into three main groups:

A. Freely available data

B. Data with limited access due to rights considerations

C. Data with limited access due to confidentiality and privacy considerations

A Freely available data

The Government's AI strategy has as a "goal that data that can be made openly available shall be shared so that they can be used by others." [126]

There are a number of datasets and sources that can freely be used for training language models, for example websites from public health authorities:

- Health professionally quality-assured knowledge and information, for example from NIPH, specialist health services, Helsenorge.no, Helsebiblioteket.
- National requirements and recommendations from the Norwegian Directorate of Health

- Method books from specialist health services
- Free professional language resources, for example health professional terminologies and classifications without license requirements, for example future ICD-11 and terms in SNOMED CT

B Data with limited access due to rights considerations

Some data may have limited access due to rights considerations, such as copyright, for example:

- Professional and textbooks (See "Mímir project" below)
- Knowledge bases and reference works, for example Felleskatalogen and Store medisinske leksikon
- Professional language resources, for example health professional terminologies and professional dictionaries with license requirements

Mímir project and training on copyright-protected material

On behalf of the Government, the National Library collaborated with NorwAI at NTNU and Language Technology Group at UiO in 2024 to analyze the significance of copyright-protected material on linguistic quality in language models.

The language in language models with and without copyright-protected material such as Norwegian newspapers and books was compared. The results showed that language models trained on content where rights-protected Norwegian material is included achieve better quality.

The goal of the project was to gather empirical evidence that could form the basis for possible agreements between the state and rights holders on the use of copyrighted content for AI purposes. Work is now underway to establish principles related to such agreements (as of March 1, 2025).

Source: https://www.nb.no/content/uploads/2024/08/Mimirprosjektet_teknisk-rapport.pdf

Training on professional language resources

There are several professional language resources that can be used to train language models, including terminologies and classifications.

SNOMED CT is an international machine-readable terminology with nearly 370,000 health professional concepts. Approximately 130,000 of the concepts are translated into Norwegian and cover health professional areas such as anatomy, findings, symptoms, diagnoses, procedures, substances, and medicines. SNOMED CT is mainly used for documentation and interaction of patient information.

The translation is done in Bokmål, while an increasing part is also available in Nynorsk. The resource is multilingual in that there is a direct connection between English and Norwegian terms.

The Norwegian translations are freely available from the Language Bank at the National Library. Internationally, the terms in SNOMED CT have been used to train several language models (pubmed.ncbi.nlm.nih.gov).

The Norwegian Directorate of Health manages the Norwegian version of SNOMED CT with regular upgrades.

SNOMED CT also contains a deeper machine-readable structure (ontology) with, among other things, concept relationships. SNOMED CT's ontology has been used for RAG in language models, also in Norway, but requires a license from the Norwegian Directorate of Health.

Another relevant resource is the International Statistical Classification of Diseases and Related Health Problems (ICD), which is owned and managed by WHO. The health service in Norway currently uses ICD-10 as a code system for diseases and causes of death and therefore contains terms and concepts that are in established use internationally.

The Norwegian Directorate of Health manages the Norwegian version of ICD. WHO has now updated ICD-10 to ICD-11. ICD-11 has updated medical content that is much more comprehensive and also includes a terminology. The resource is now being translated into Norwegian and will be freely available for use.

Similarly, ICPC-2 is the international classification for health problems, diagnoses, and other reasons for contact with primary health services. ICPC-2 is in use in Norway and the Norwegian Directorate of Health maintains both this and the website where the updated international edition of ICPC-2 (English version, ICPC-2e-English) is published on behalf of the Wonca International Classification Committee (WICC).

The Norwegian Directorate of Health also manages a number of other relevant classifications, for example the Norwegian Procedure Code System, Norwegian Laboratory Code System with associated code systems (Test Material, Anatomical Localization, Textual Result Values, and Examination Method), Norwegian Pathology Code System (NORPAT), and Activity Codes for pathology laboratories (APAT). The development of the Norwegian Procedure Code System was initiated by the Nordic Council and still has a Nordic-Baltic core (NCSP). This contains, for example, terms for procedures and procedure groups in use in specialist health services in the Nordic-Baltic countries.

C Data with limited access due to confidentiality and privacy considerations

Several data sources have limited access due to confidentiality and privacy considerations. This applies, for example, to:

- Texts from patient records (electronic health records)
- Data from quality registries and other health registries
- Health surveys such as HUNT
- Applications and responses related to Norwegian health administrative businesses

Training on text from patient records, Clinical NorBERT

Helse Vest IKT has, in collaboration with the health enterprises in Western Norway, trained a language model using, among other things, journal texts. To safeguard privacy, the texts have been anonymized by replacing place names with another place name, first and last names with another first and last name, dates with another date, etc. As part of the research project, the quality of the anonymization has been analyzed.

To gain access to the journal texts for training the model, an application has been made and granted exemption from confidentiality obligations for research purposes in the Health Personnel Act by the Regional Ethics Committee (REC).

It is envisioned that Clinical NorBERT can be used for, for example, automated text analyses and machine-assisted coding. The language model is a BERT model that differs from generative language models. Therefore, there is little risk that the raw data on which the language model was originally trained can be recreated.

Clinical NorBERT can be made available for use by other actors in health and care services under certain conditions. If the language model is to be fine-tuned for a specific area of use, new approval from REC or the Norwegian Directorate of Health is required to use health data for this purpose.

Helse Vest IKT is now looking at possibilities for training generative language models.

Source: Helse Vest IKT

Data sources such as texts from patient records can be particularly relevant for training large language models because they reflect actual language use and healthcare personnel's use of knowledge. However, it can be time-consuming to access such data [127]. This is partly because healthcare personnel are subject to confidentiality obligations, as per Chapter 5 of the Healthcare Personnel Act. This limits free processing of health information from patient records and other health registries for training and use of language models. Section 29 of the Healthcare Personnel Act nevertheless opens for dispensation from confidentiality obligations for use of health information from treatment-oriented health registries (patient records) when certain conditions are met [128]. An example of such dispensation being granted is texts from patient records for training the language model Clinical NorBERT at Helse Vest IKT and NorDeClin-BERT at the National Center for E-health Research.

Synthetic data

Limited access to sensitive personal data has led to discussion about the need to create synthetic texts [129]. Synthetic data that cannot be traced back to individuals is not personal data and is therefore not covered by confidentiality obligations. However, there are also a range of challenges related to synthetic data.

Research on ASR systems (Automatic Speech Recognition) and downstream AI models shows that synthetic data can have significant limitations. Research shows that synthetic transcriptions can contain errors, hallucinations, and unnatural language structures that affect the performance of models using these data [130]. For example, a study shows that using simulated ASR output to train models (by using text-to-speech and then speech-to-text) could make models more robust against errors, but quality still varied compared to authentic data [131].

A particular concern with using synthetic health data is that hallucinations in language models can lead to false information being incorporated into datasets [132]. One study shows that even advanced systems like Whisper can "invent or 'hallucinate' entire phrases and sentences" in about 1% of transcriptions [133]- In health contexts, such erroneous additions can lead to models being trained on medical information that was never spoken, which can have serious consequences for diagnosis and treatment recommendations.

In addition, one is, as mentioned, dependent on authentic data to create synthetic data, so the problem related to privacy can not necessarily be avoided. The quality of synthetic data depends heavily on the quality and representativeness of source data, and the study indicates that certain patterns of error in original data can be amplified or create new problems in synthetic datasets.

There is therefore still a need for more knowledge to be able to say whether synthetic datasets can be sufficiently suitable for training language models, particularly when it comes high risk domains such as the health sector. Researchers recommend a combination of improved ASR technology, error-correcting methods, robust training, and cross-modal validation to reduce problems with synthetic data, but these challenges are still not fully solved [134].

Infrastructure for computational resources

Adaptation of language models to Norwegian conditions will require extensive computational power and associated infrastructure. The need will depend on many factors, including how the adaptation is carried out.

The Research Council of Norway has conducted a concept selection study on the need for computing power and organization of national infrastructure [135]. The report points to an investment need of 3.4 billion kroner over the next five years. A cost-benefit assessment for each administrative area will be conducted in step 2. However, this step does not include the health and care sector and points out that a separate study is required since the sector is extensive and complex with special requirements for privacy.

Recent technological developments suggest that much more efficient ways of developing language models are now being developed than previously, which may reduce previously assumed needs for computational power.

However, the large amount of sensitive data will still present challenges related to infrastructure for computing power. There may therefore be a need for suitable infrastructure capable of handling such data for training and fine-tuning of language models.

Quality framework for large language models

Evaluation of language models is complex and will encompass several types of tests (benchmarks). Relevant research has been conducted on tests for adapted language models to general Norwegian conditions, for example language, at the University of Oslo [136].

The evaluation should encompass both general and user-oriented testing. General quality measurement can be based on standardized tests that evaluate basic medical knowledge, for example through adapted versions of medical examination questions. User-oriented and context-specific quality measurement can be testing in real or simulated use situations that are representative of the model's intended use in Norwegian health and care services.

Testing using medical examination questions

International research often uses medical examination questions to test the quality of language models. A common dataset is MedQA, which includes over 60,000 questions from several medical examinations in the USA and China.

Such datasets can be relevant for testing general medical competence, for example anatomy and diagnoses. However, the clinical everyday life where AI tools are to function will differ significantly from an examination situation. Testing using examination questions will therefore not necessarily be sufficient to measure the quality of language models.

Another question is whether examination questions from the USA and China will work well in Norwegian context, and it is possible that separate datasets with examination questions based on Norwegian curriculum should be created.

An additional point is that several international language models may already be trained on datasets like MedQA. The dataset would then be contaminated and unsuitable for testing language models.

Currently, there is no generally accepted quality framework for evaluating language models for the health and care sector, but several point out that this is important to be able to use language models responsibly in health [137]. Such a framework will need to be developed gradually and be based on proven methods,

best practices, and standards. A coalition aimed at promoting responsible development of AI for health (Coalition for Health AI (CHAI™)) outlines, for example, a framework and emphasizes the importance of developing good tests to evaluate large language models with regard to five fundamental principles: usefulness, fairness and equal treatment, transparency, safety, and security and privacy [138].

Through work on this report, it has been proposed, among other things, to establish a layered model to evaluate and test various characteristics that are important for the Norwegian health and care sector. Below is a description of possible areas for testing and what a layered model for evaluation can entail.

Possible areas for testing

The quality framework can encompass tests of models within several areas.

Health professional language:

- medical terminology, including technical terms, abbreviations, and jargon
- Bokmål, Nynorsk, and possibly Sami in health professional context
- linguistic precision, language adapted to different target groups such as healthcare personnel and patients, including people with different cultural backgrounds

Health professional knowledge and practice:

- national professional guidelines
- established medical procedures and treatment protocols
- acute versus non-acute situations
- identification and explanation of medical relationships

Health administrative knowledge and practice:

- structure and organization of Norwegian health and care services
- administrative routines and procedures
- referral and documentation practice
- interaction between different levels in health services

Values and ethics:

- respect for patient rights
- privacy and confidentiality
- ethically defensible recommendations
- complex ethical dilemmas

Legislation:

- legislation relevant to the health and care sector
- legal requirements for patient treatment
- healthcare personnel's duties and responsibilities
- documentation requirements and reporting obligations

For all these areas, it is essential to establish both quantitative and qualitative methods [139].

The evaluation can include various dimensions:

- systematic evaluation of the model's accuracy and reliability

- assessment of the model's ability to acknowledge its own uncertainty
- testing the model's robustness under different conditions
- continuous evaluation of the model's performance over time to detect any performance degradation

1. Layered model for testing language models

To ensure thorough and systematic evaluation, a layered model has been proposed through the work with this report, encompassing evaluation of basic, domain-specific, use-case-specific, and context-specific properties (see layered model figure 7).

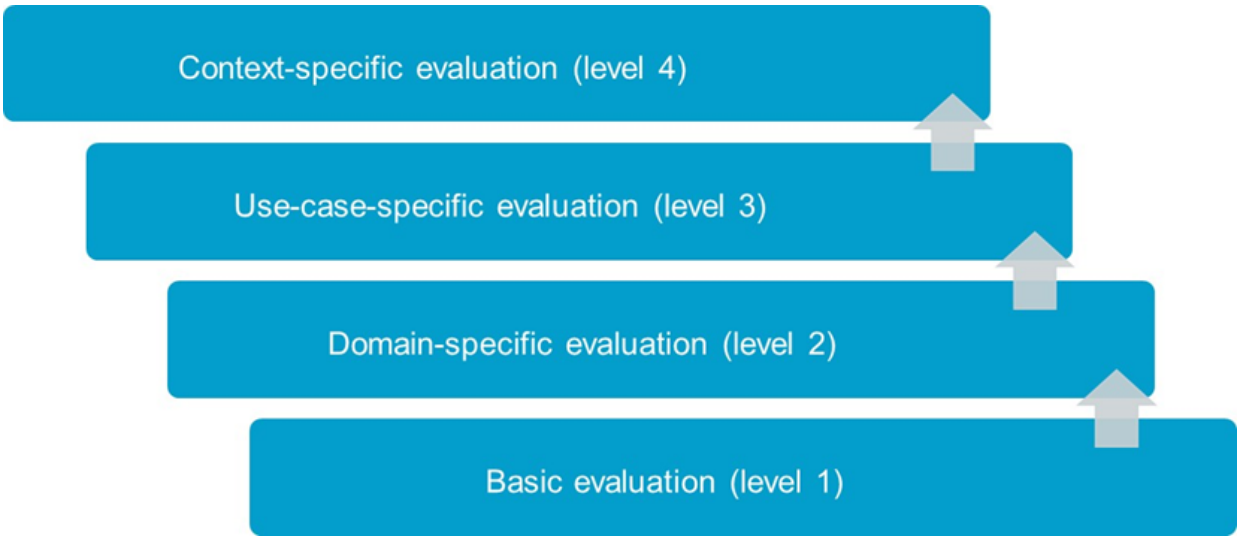

Figure 7: Layered model for evaluation and testing of language models, from basic, domain-specific, use-case-specific to context-specific

It may be appropriate to start by developing the framework for the lower levels, as well as use-cases with low risk. Where the use case falls under regulations for medical devices and/or the AI Act, standards, both existing and under development, will constitute separate frameworks that can be used to meet legal requirements.

The description below is a sketch illustrating what can typically be tested at each level and can form a starting point for further specification.

1. Basic evaluation (level 1) encompasses fundamental properties that are critical for applications in the health and care sector. This includes testing linguistic quality in both Bokmål and Nynorsk and Sami where relevant, and how the model handles general medical language and terminology, basic security and privacy, as well as technical performance such as response time and stability. At this level, it can be considered to establish minimum requirements that all models must meet to be used in the health sector.
2. Domain-specific evaluation (level 2) encompasses domain-specific properties that are important for the health and care sector. This can include testing the model for specialized medical professional language, clinical guidelines, health administrative processes, and ethical frameworks. At this level, the model's ability to handle regional and local conditions in the Norwegian health and care sector is also assessed.
3. Use-case-specific evaluation (level 3) encompasses specific use cases within the health and care sector. For example, a model to be used in emergency care would be tested specifically for the ability to assist with triage assessments, emergency medical procedures, and coordination with other departments. A model for creating speech-to-summary of patient conversations should be tested for exactly this use case. Other use cases will have other specific requirements that are evaluated. Testing at this level helps assess how suitable the model is for its intended purpose.
4. Context-specific evaluation (level 4) concerns the model's suitability in the specific implementation context. This encompasses testing integration with local systems, adaptation to established work

processes, and handling of specific documentation and privacy requirements. Testing at this level also assesses the model's ability to meet local quality indicators and specific needs in the implementation environment.

Some advantages of a commonly accepted testing framework are that it:

- provides a basis for comparison between different models
- enables systematic evaluation from the general to the specific
- simplifies identification of weaknesses and areas for improvement
- facilitates targeted optimization
- ensures that critical aspects are evaluated
- streamlines the testing process by revealing fundamental shortcomings early

Testing should not only be done once, but continuously over time, to capture changes in performance, identify new challenges, and ensure lasting quality in the service. The framework should be regularly revised and updated in line with technological development and new experiences from practical use.

Competence in development and use

It is crucial that the health and care sector has sufficient access to necessary competence throughout the entire lifecycle of language models, from development, adaptation, testing, and use of language models adapted to Norwegian conditions [140]. Several informants have provided feedback that more competence in the health and care sector is necessary, including in artificial intelligence (AI), information and communication technology (ICT), health sciences, law, linguistics, and economics.

For the use of AI tools as part of healthcare, the requirements that follow from health legislation apply as elsewhere. The health and care service has a duty to ensure that the healthcare provided is sound. Among other things, it must ensure that the AI tools used contribute to making the healthcare provided safe and secure. According to the AI Act, there is a responsibility on organizations that deploy AI solutions to train users. The organization also is also responsible for delegating tasks related to ensuring human oversight to persons with the necessary competence and who have undergone the necessary training to perform this function [141].

AI literacy is a central concept in the AI Act. There are also attempts to operationalize and specify the concept at the levels of 'knowledge', 'understanding', and 'skills' [142].

Research is also being conducted on AI literacy. An international meta-study indicates that healthcare personnel and students have low AI literacy [143]. We are not aware of a comprehensive overview of Norwegian conditions, but it may be reasonable to assume that there is insufficient AI literacy for safe and effective use, development, and testing of language models in Norwegian health and care services.

Governance of language models

Today there is no national governance, including infrastructure, for large language models. The technology is still rapidly evolving, and new international language models are constantly being introduced. Language models are also being developed and adapted to Norwegian conditions by actors in both the public sector and industry. However, there is no comprehensive national overview of which models exist and which are used in the health and care sector.

A stable governance structure can facilitate safe introduction of language models in health and care services by ensuring quality, relevance, and legal frameworks, among other things.

The Technology Council recommends to define selected Norwegian language models as a national common service, similar to ID-porten, Altinn, and the Population Register, to ensure good operation, management, and access [144]. Both Norwegian pre-trained language models and adapted language models can be included. The Technology Council further points out that a new function should be established for development and governance of such a common service. This could also apply to the health and care sector, like other common services within the sector. National management could address several of the risks described above through, for example, testing and evaluation of language models for the sector.

Common governance could encompass several functions, including:

- governance and further development common data resources for adaptation to Norwegian conditions
- governance and further development of quality frameworks
- governance of common language models
- regulatory guidance for the health and care sector
- collaboration with research institutions and participation in relevant research projects
- making language models available for use in the sector or further development by, for example, suppliers or the public sector
- coordinating efforts in the sector related to language models

Common governance would not replace industry's role as supplier, but rather facilitate safe frameworks for innovation and use, for private and public organizations. Further development by commercial actors can contribute to a broader range of solutions. For example, a governance organization can test, adapt, and make available one or more pre-trained healthcare language models. Such models can be further developed for specific use cases by, for example, private or public actors.

A governance structure can ensure long-term sustainability, quality, and coordinated development of language models for the sector. The organization should be responsible for following technological development, assessing new opportunities, but also risks, and establishing clear frameworks for how models can be applied, with particular emphasis on patient safety and privacy. This way, it can provide advice on implementation and use of language models in the health and care sector.

Governance can take place in an existing organization, or a separate organization can be established, for example, a dedicated center. There will be a need for closer investigation of needs for and how such management should be organized.

[122] https://www.regjeringen.no/en/dokumenter/the-digital-norway-of-the-future/id3054645/ p.70

[123] https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00224-3/fulltext

[124] https://code-of-practice.ai/?section=transparency#model-documentation-form

[125] https://huggingface.co/docs/hub/en/model-cards

[126] https://www.regjeringen.no/contentassets/1febbbb2c4fd4b7d92c67ddd353b6ae8/en-gb/pdfs/ki-strategi_en.

[127] https://ehealthresearch.no/nyheter/2024/en-revolusjon-for-helsesektoren-den-forste-norske-kliniske-sprakm

[128] https://www.helsedirektoratet.no/rundskriv/helsepersonelloven-med-kommentarer/taushetsplikt-og-opplysni

[129] https://tidsskriftet.no/2024/11/kronikk/syntetiske-datasett-kan-gi-bedre-ki-modeller-helsetjenesten

[130] https://arxiv.org/html/2408.14418v1, https://dl.acm.org/doi/abs/10.1145/3630106.3658996

[131] https://arxiv.org/abs/2108.13048

[132] https://arxiv.org/html/2401.01572v1

[133] https://dl.acm.org/doi/abs/10.1145/3630106.3658996

[134] https://arxiv.org/html/2412.00055v1

[135] https://www.forskningsradet.no/nyheter/2025/investere-milliarder-infrastruktur-tungregning/

[136] https://arxiv.org/abs/2412.06484

[137] https://jamanetwork.com/journals/jama/fullarticle/2825147

[138] https://chai.org/generative-ai-work-group/

[139] Examples of dimensions and axis described in https://jamanetwork.com/journals/jama/fullarticle/2825147 and https://arxiv.org/abs/2211.09110

[140] https://healthinformaticsjournal.com/downloads/files/524.pdf

[141] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[142] https://www.pwc.nl/en/services/audit-assurance/pwc-accountancy-insights/data-it-and-internal-control/ai-lite

[143] https://healthinformaticsjournal.com/downloads/files/524.pdf

[144] https://teknologiradet.no/publication/generativ-kunstig-intelligens-i-norge/

Who Can Adapt to Norwegian Conditions?

Several actors are relevant for adapting language models to Norwegian conditions.

The university and higher education sector

In the university and higher education sector, there are several initiatives that both develop pre-trained Norwegian language models and fine-tune international models, for example, the Norwegian Research Center for AI Innovation (NorwAI) at NTNU. The NORA consortium includes many collaboration partners, including UiO and Norwegian Computing Center.

The National Center for E-health Research developed a healthcare language model (NorDeClin-BERT).

Future research projects related to language models will play an important role in adaptation to Norwegian conditions. It is important that some of these initiatives include adaptation to the Norwegian health and care sector, i.e., healthcare language models.

The health and care sector

In the health and care sector, both Helse Vest IKT and Sørlandet Hospital have been involved in adapting language models to Norwegian conditions for research purposes. Helse Vest IKT is now also looking at specific use cases for the language model Clinical NorBERT.

Industry plays a very important role by fine-tuning language models and offering services based on language models, including Norwegian pre-trained models, but especially international models. Suppliers now offer services based on language models to, among others, general practitioners.

The National Library

The National Library is a central research institution that develops and adapts language models for Norwegian conditions. The National Library has developed and made available the language model NB-Whisper for transcription of audio recordings, and the model has been further developed by suppliers for the health and care sector.

The National Library will most likely have a special position in Norway when it comes to adapting language models to Norwegian conditions. The government proposes in the state budget for 2025 to allocate a total of 40 million kroner for training Norwegian and Sami language models (foundation models) for use in artificial intelligence. Of these, 20 million kroner will go to the National Library, which is tasked with training and making language models available, while 20 million kroner will go to financing computing power for training the models. The computing power will be delivered by the high-performance computing company Sigma2 AS, which is owned by Sikt – The Knowledge Sector's Service Provider [145]. The National Library's position could become very important for work with language models in the health and care sector as well, and close collaboration could be appropriate.

[145] https://www.regjeringen.no/contentassets/e628c0d8c3a44971bb48bd5a487cb52d/nn-no/pdfs/prp20242025

Recommendations

For safe and secure use of AI tools in health and care services, it is crucial that the AI models and systems are of high quality. Being able to adapt large language models to Norwegian health and care services could help reduce several risks associated with using AI tools in healthcare.

The government's digitalization strategy has AI as a separate focus area and will, among other things: [146]

- work to further develop a national infrastructure for AI that will provide access to foundation models based on Norwegian and Sami languages and social conditions
- investigate how data and text from public organizations can be used for ethical and safe training of national language models
- clarify what constitutes lawful use of intellectual property and other protected works in text and data mining processes

In the State Budget for 2025, 20 million kroner is allocated to the National Library's work with artificial intelligence and training of generative language models [147]. The Research Council of Norway has allocated over 1 billion kroner to establish four to six AI research centers across sectors to address challenges such as societal challenges from digital technology development [148].

The government announced on March 21, 2025, that they would establish "AI Norway," which will be a new national arena for innovative and responsible AI and part of the national administrative apparatus now being established. At the same time, the government announced that the Norwegian Communications Authority (Nkom) will have supervisory responsibility for the EU's AI Act [149].

The recommendations in this report related to adaptations to the Norwegian health and care service will, to the greatest extent possible, seek synergies and collaboration with relevant work and stakeholders at the national level.

To facilitate the use of generative AI models based on foundation models adapted to conditions in the Norwegian health and care service, the Norwegian Directorate of Health recommends the following measures:

1. establish quality frameworks for large language models used in Norwegian health and care services
2. build and share competence on development and use of large language models
3. establish common data foundation
4. investigate infrastructure with computing power adapted to the health and care sector's needs
5. investigate a governance structure for large language models for the health and care sector

Most measures will have to be solved in collaboration between a number of actors, both in the health and care sector and outside, such as the National Library, the university and college sector, the Digitalization Directorate, and the Research Council.

[146] https://www.regjeringen.no/en/dokumenter/the-digital-norway-of-the-future/id3054645/

[147] https://www.nb.no/pressemeldinger/regjeringen-vil-satse-stort-pa-digitalisering-og-kunstig-intelligens-ved-nasjon

[149] https://www.forskningsradet.no/forskningspolitikk-strategi/ltp/kunstig-intelligens/

[150] https://www.regjeringen.no/en/aktuelt/gjor-norge-klar-for-trygg-og-innovativ-ki-bruk/id3093081/

Establish quality framework for large language models for Norwegian health and care services

Foreslått tiltakseier: Helsedirektoratet i første omgang

Proposed responsible party: Norwegian Directorate of Health, initially

In collaboration with: Health and care services, Norwegian Institute of Public Health, Health Supervision Authority, DSA, R&D environments, National Library, Digitalization Directorate, Research Council of Norway, Innovation Norway, etc.

Relevant for: The health and care sector

Problem to be solved: There is great uncertainty about what constitutes sufficiently good quality to use large language models in the health and care sector, and how this is measured. Evaluation of language models is complex and there are currently few generally accepted methods for evaluation and testing (benchmarks) of large language models for the health and care sector. This also makes it difficult to choose the right language model for further development into high-quality AI systems.

Proposal: The sector develops and establishes common quality frameworks for testing and evaluating language models to contribute to safe use of generative AI models in the health and care sector.

The establishment of the quality framework will require close collaboration with several stakeholders in the health and care sector, public and private sector, research sector, and business.

A good starting point would be administrative use cases with low risk and with basic evaluation, which is the lowest level in Figure 7. The framework can be gradually expanded to higher-risk use cases as the field matures and the EU clarifies both regulations and standards related to AI and medical devices.

The work should include the following:

1. Identify organizations that should participate in the work and clarify roles and responsibilities
2. Map relevant existing tests (benchmarks), frameworks, best practices, and standards for testing and evaluation of large language models
3. Systematize experiences from similar work in the sector, including the Norwegian Directorate of Health's work with AI services such as Helsesvar and Easier Access to Information (ETI).
4. Gradually develop and pilot the framework based on selected use cases

The framework can include both quantitative and qualitative evaluation methods. For further details, see chapter 4.3 and chapter 4.4.

Build and share competence on development and use of large language models

Proposed responsible party: Several measures, with different responsible parties

In collaboration with: Norwegian Directorate of Health, the health and care sector, other relevant environments, such as National Library, Digitalization Directorate, R&D institutions and higher education institutions.

Relevant for: The health and care sector

Problem to be solved: There is a lack of necessary competence to adapt, evaluate, and implement language models in the Norwegian health and care service.

Proposal: The health and care sector works systematically to build and share competence, especially through experiences from practical testing and use.

The work should include:

1. This report can be used to raise competence and increase awareness of risks and risk-reducing measures, in health services, internally in the Norwegian Directorate of Health, in health authorities, and throughout the sector. Dissemination can be done actively through presentations, webinars, etc.
2. Further develop cross-agency regulatory guidance service with topics related to the use of large language models in health and care services.
3. Follow up seminar with general practitioners, related to the use of AI tools in general practice, in collaboration with the Norwegian Association for General Practice.
4. Work to ensure that ongoing and future research and development projects on AI also include language models adapted to the Norwegian health and care sector.
5. Follow Nordic and international approaches to development, testing, and use of language models in the health and care sector
6. Share experiences and competence through, for example, seminars, reports, and guidance materials
7. Pilot adaptations of large language models to the Norwegian health and care service in collaboration with one or more key actors, for instance the National Library.
8. Pilot frameworks for testing/quality measurement of language models in real use situations in close collaboration with professional environments that can contribute to adaptations to Norwegian health and care services

For further details, see chapter 4.3 and chapter 4.4

Establish shared data resources

Proposed responsible party: Several measures, with different responsible parties

In collaboration with: Norwegian Directorate of Health, the health and care sector, regional health enterprises, Norwegian Association of Local and Regional Authorities (KS), NIPH and National Library, DSA

Relevant for: the health and care sector

Problem to be solved: The health and care sector lacks a sufficient common foundation of data to develop, adapt, and test language models more easily and cost-effectively, and thus contribute to the development of high-quality AI tools adapted to Norwegian conditions.

Proposal: The health and care sector should collaborate to make more high-quality data available for both training (pre-training and post-training), knowledge grounding (e.g., RAG), and testing of language models to be used in the health and care sector.

The work should include:

1. mapping open and easily accessible data sources, such as guidelines and methodology books, for developing large language models for health and care services (Hdir, in collaboration with NIPH). An overview of freely available texts and other language resources with healthcare knowledge and practice can be placed on the information page about AI at the Norwegian Directorate of Health
2. making professional language data sources such as terminologies and classifications available for reuse together with other relevant data, for example, on helsedata.no, helsedirektoratet.no, or in the language bank at the National Library (Hdir), also Nynorsk and Sami
3. establishing data resources that constitutes common national principles for values and ethics in language models in Norwegian health and care services
4. assessing how sensitive data such as medical records, discharge summaries, and forms can be used to develop large language models (clarify regulations through guidance, technical solutions, etc.) (Hdir, in collaboration with relevant (research) environments)
5. making data sources protected by specific rights more accessible, for example, by considering establishing a compensation scheme for rights holders of healthcare content in line with national principles (see the Mimir project at the National Library [150])
6. investigating the need for and establishment of a common infrastructure for sharing language data, for example, through helsedata.no (NIPH, Hdir, health enterprises). See also recommended measures under establishing infrastructure for computing power (chapter 5.2)

For further details, see chapter 4.3 and chapter 4.4

[150] https://www.nb.no/content/uploads/2024/08/Mimirprosjektet_teknisk-rapport.pdf

Investigate infrastructure with computational power adapted to the health and care sector's needs

Proposed responsible party: Several measures, with different responsible parties

In collaboration with: Norwegian Directorate of Health, Norwegian Health Network (NHN), Research Council, National Library, the university and college sector, regional health enterprises, regional ICT companies and Norwegian Association of Local and Regional Authorities (KS), the health and care sector in general

Relevant for: the health and care sector

Problem to be solved: The health and care sector is a comprehensive and complex sector with special requirements for privacy. Existing infrastructure for computational power in the sector is largely unsuitable

for handling large amounts of sensitive data. However, there are no plans to investigate the need for computing power for developing and using large language models specifically for the health and care sector.

Proposal:

An investigation is initiated to assess needs, costs, and benefits related to computing power and infrastructure to be able to train, ground, test, and use large amounts of sensitive data in the health and care sector and propose different approaches to solve the need.

For further details, see chapter 4.3. and chapter 4.4. Some of the need for such infrastructure is described in chapter 4.3.4 in the Joint AI plan for the safe and effective use of AI in the Norwegian health and care services [151].

[151] https://www.helsedirektoratet.no/rapporter/joint-ai-plan-for-the-safe-and-effective-use-of-ai-in-the-norwegian

Investigate a governance structure for large language models in the health and care sector

Proposed responsible party: Several measures, with different responsible parties

In collaboration with: Norwegian Directorate of Health, regional health enterprises, Norwegian Association of Local and Regional Authorities (KS), regional ICT companies, National Library, Norwegian Communications Authority, Digitalization Directorate, Ministry of Digitalization and Public Governance, Ministry of Education and Research

Relevant for: the health and care sector

Problem to be solved: Lack of governance of language models is a barrier to efficient use of resources and sharing of competence in the field. Little coordination can result in varying and poor quality of language models in the sector. Language models are currently shared to a limited extent between actors in the sector.

Proposal: Investigate the organization of a separate organization/center responsible for governing language models in the health and care sector.

A governance structure could encompass several functions, including

- governance and further development common data resources for adaptation to Norwegian conditions
- governance and further development of quality frameworks
- governance of common language models
- regulatory guidance for the health and care sector
- collaboration with research institutions and participation in relevant research projects
- making language models available for use in the sector or further development by, for example, suppliers or the public sector
- coordinating efforts in the sector related to language models

> For further details, see chapter 4.3. and chapter 4.4.