Policy on the Use of Generative AI for Graduate Studies

Department of Pathology & Laboratory Medicine

Western University

1. Introduction

Artificial intelligence (AI) is the study of enabling a computer to perform tasks that would otherwise require a human intelligence. This field of research has existed for many decades. However, the rapid emergence and widespread popularity of large language models (LLMs, colloquially known as 'AI chatbots') has caused the term 'AI' to become synonymous with this latest iteration of AI models. An LLM is a deep neural network model, consisting of billions of parameters, that is designed to generate text in response to a prompt. For a more detailed introduction to the history of AI and the development of LLMs, please see Appendix 1: Understanding Generative AI.

The purpose of this document is to:

- review the current guidelines for AI usage at Western University, and the School of Graduate and Postdoctoral Studies (SGPS) within Western;
- provide an overview of the ethical and scientific issues arising from the use of generative AI;
- and describe the Department's policy for AI usage for graduate research.

A summary of the key points is provided at the top of most sections or subsections for your convenience. However, you are expected to read through the entire contents of each section.

This document was prepared with the LaTeX typesetting engine. The initial draft was written by Art Poon, Chair of Research-Based Graduate Education for the Department of Pathology & Laboratory Medicine (PaLM). Drafts were reviewed by members of the PaLM Graduate Education Committee.
First version: June 15, 2026.
Second version (SGPS updates Thesis Procedures): June 22, 2026.
Third version (Add policy on GTA, supervisor responsibilities): June 23, 2026.

2. University Guidelines

2.1. Principles of Using AI

- Do not use AI to replace your critical thinking and expertise.
- Meaningful AI usage must be clearly disclosed.
- Do not submit personal or sensitive information to public AI tools.
- You are accountable for the accuracy of AI-generated outputs.
- Be mindful of biases in AI outputs.
- Methods of using AI should be transparent and reproducible.

This section reproduces the text published at *https://ai.uwo.ca/using-ai/principles-of-using-ai.html* (accessed June 5, 2026).

Experiment responsibly, safely, and ethically. Use AI to support your work — brainstorm ideas, draft early versions, organize information, summarize notes, or improve clarity. But don't use it to replace your own thinking, judgment, or creativity, especially for tasks that require ethical reflection, nuance, or expertise.

Here are five principles to guide our community's engagement with generative AI:

- **Integrity:** The use of generative AI in academic work must be clearly disclosed. Cite or acknowledge AI when it plays a meaningful role in your work, whether in documents, presentations, research, or image generation.
- **Privacy:** Personal data must be adequately protected. Never input confidential, personal, or sensitive information into public AI tools. If you're collecting or working with personal data, visit Western's Legal Counsel website to learn about Privacy Impact Assessments.
- **Accountability:** You bear responsibility for the consequences of AI-generated outputs. Always review and fact-check what AI produces. Never assume it's accurate, unbiased, or appropriate for your context.
- **Inclusion:** Actively consider accessibility and fairness. AI tools can reflect and amplify existing biases. Be mindful of who might be affected by the output.
- **Transparency:** The algorithms, data, and design decisions underlying AI systems should be openly accessible to the extent possible. Be open about when and how you're using these tools.

2.2. Guidance for use in graduate studies

- Act with academic integrity.
- Adhere to program expectations and course syllabi.
- Discuss AI usage with your supervisor and advisory committee.
- Follow conference and journal policies on disclosure of generative AI use.

This section reproduces the guidance document published by the School of Graduate and Postdoctoral Studies (SGPS; last accessed June 24, 2026). The original document can be accessed at *https://grad.uwo.ca/about_us/policies_procedures_regulations/ai.html*.

With ongoing advancements in artificial intelligence (AI), the School of Graduate and Postdoctoral Studies recognizes the need for guidance on the ethical and responsible use of generative AI technologies in graduate work.

You have an obligation to act with academic integrity and abide by the syllabus for each course, and the expectations for each milestone, including the thesis or major project, within your graduate degree program. This obligation applies regardless of the physical or virtual location of graduate program activities. Where you are uncertain, ask your graduate program chair (or equivalent) for guidance.

Guidance for Use: In performing activities that are part of the degree requirements (for example, course work, milestones, and thesis research), graduate students may fall under the categories of both student and researcher as outlined in the Guidance by Role for Western University.

Students: For course work, the use of generative AI must abide by the course learning activities and policies, as outlined in the syllabus. Where you are uncertain, ask your course instructor for guidance.

For research, scholarship, and creative activities, the use of generative AI should be discussed with the supervisor(s) and supervisory committee. Any plan for the use of generative AI must follow the Principles of Use outlined above and must be reviewed as necessary to ensure ongoing adherence to these principles. Description and disclosure of the use of generative AI in research methods, scholarship, and creative activities should be consistent with the requirements of the discipline(s) or field(s) of study. Where you are uncertain, ask your supervisor(s), supervisory committee, or graduate program chair (or equivalent) for guidance.

Researchers: The use of AI in primary research, scholarship, and creative activities is governed by the same policies and regulations that govern non-AI-assisted research at Western. Graduate student researchers are expected to abide by the policies governing responsible conduct of research at Western. Researchers must also abide by any policies or guidance regarding the use, and disclosure of use, of generative AI provided by the agencies, organizations, foundations, and companies which fund their research. This requirement also applies to graduate scholarship and award applications. When disseminating their findings, researchers must abide by the policies of event organizers and publishers regarding disclosure of generative AI use.

Effective July 1, 2026, SGPS has updated the Procedure for Thesis Formats and Content with the following addition:

4.7 - Statement on the use/non-use of Generative Artificial Intelligence

- This section must include a high-level overview of the use or non-use of Generative Artificial Intelligence (GenAI) in the thesis. Non-use of AI should be stated explicitly.
- The use of GenAI must not undermine the integrity of the research and the thesis including the originality of the student's work.
- In addition to this high-level overview, any use of GenAI in the generation, interpretation, or presentation of research content or knowledge must be appropriately documented or acknowledged throughout the thesis in accordance with disciplinary standards.

You are advised to check the websites linked above for up-to-date information on new and revised policies from the University and SGPS.

3. *Pitfalls and ethical issues in AI use*

- Large language models (LLMs) are susceptible to generating false and misleading information ('AI hallucinations').
- Widespread use of AI tools is fuelling a crisis in the peer review of grants and scientific manuscripts.
- LLMs are prone to amplifying systemic biases in training data.
- Data centres deploying LLMs at scale consume significant amounts of energy and freshwater.
- Over-reliance on AIs has known risks to your ability to learn and your mental health.

There are a large number of issues that have arisen around the use of generative AI in different settings, including its environmental impact, risks to your mental health, and susceptibility to producing false and misleading information ('AI hallucinations'). Commercial generative AI models have also been refined by outsourcing reinforcement learning tasks to human labour in the Global South, representing a hidden human cost to the development of these tools[1]. Many people have chosen to avoid any use of AI in any circumstance for these reasons. The following list is not a comprehensive overview of all issues, but attempts to focus on issues that pertain to graduate research, and why you may choose to avoid using AI for specific tasks, to reduce your dependence on these tools in general, or to reject them altogether.

3.1. Mental health

Graduate studies can create an environment where students may experience high levels of stress and isolation. Evans et al., 2018 reported that "graduate students are more than six times more likely [41%] to experience depression and anxiety as compared to the general population". There have been numerous reports of mental health crises associated with the excessive use of generative AI 'chatbots' (Dohnány et al., 2026). For example, AI psychosis is a recently documented phenomenon that occurs when a person develops delusions or paranoia from extensive interactions with generative AI chatbots (Flathers, Roux, and Torous, 2026). A recent study used a set of structured scenarios to evaluate whether AI chatbots exacerbated or intervened on the development of delusional beliefs in simulations of vulnerable users (Yeung et al., 2025). They observed that the chatbots tended to

[1] An investigative journalism piece was recently published by TIME Magazine, reporting that workers in Kenya, Uganda and India were paid "between $1.32 and $2 per hour" to review and label "tens of thousands of snippets of text" to improve OpenAI's ChatGPT model. Some of these texts "described situations in graphic detail like child sexual abuse, bestiality, murder, suicide, torture, self harm, and incest" (Perrigo, 2023).

reinforce rather than challenge user delusions, and attempted to prevent user self-harm in only about one-third of interactions. Always remember that generative AI chatbots are commercial products. They are intentionally designed to retain your engagement (*e.g.*, by expressing praise and excessive obedience) and to minimize the chance that you will switch to a competing product.

3.2. Inaccurate outputs

A well-known problem in the use of LLMs is that the models can generate text that is misleading or false, despite being grammatically correct. This can happen if the user submits an ambiguous or incorrectly-worded prompt, or if the training data are inaccurate or outdated. Even with perfect data and a well-composed prompt, however, an LLM can nevertheless generate false information — these instances are sometimes referred to as AI hallucinations. In the context of digital pathology, hallucinations have become an issue when generative AI models are used to augment histopathology slide images (known as 'virtual staining'), where the model fabricates image features that cause the pathologist to misclassify the sample (L. Huang et al., 2025).

AI models can produce misleading outputs without hallucination when the models are insufficiently validated or trained on non-representative data. For example, the Epic Sepsis Model is a proprietary AI-based tool for predicting sepsis in hospital in-patients. It had been deployed at hundreds of hospitals until external validation studies determined that this tool had a 67% false negative rate (failing to detect actual cases of sepsis) (Wong et al., 2021). The system also generated alerts on 18% of the general hospital population (a high false positive rate); the resulting alert fatigue drove some floor nurses to disable the tool by "covering the video camera" (J. W. Gichoya et al., 2023). This model was eventually withdrawn and revised by the vendor. The well-documented failure of the Epic tool has been attributed to data leakage during training (the training data included orders for antibiotics), as well as ethnic and socioeconomic differences between the populations represented by the training dataset and the healthcare settings in which the system was deployed[2].

3.3. Misuse of generative AI in science

The release of increasingly powerful generative AI tools has made it feasible to produce images, text and other media that would normally require sufficient time, resources, and the development of specific skills (*e.g.*, watercolour painting, playing the piano). Instead, these media can be generated by non-expert users with minimal effort. However, the resulting media tend to be perceived as 'low quality'

[2]The Epic model was initially validated at the University of Colorado Health system that serves a predominantly white population. An external validation study from Ostermayer et al., 2024 set in Harris County, Texas, with a predominantly Hispanic and Black population, reported a sensitivity of 14.7% and noted the known differences in sepsis among races.

and are prone to artifacts induced by AI hallucinations due to insufficient validation by the user — it has become common to pejoratively refer to such outputs as 'AI slop'. These outputs tend to be highly homogeneous, which can also be attributed to minimal effort to compose and customize prompts. Generative AI models tend to 'regress to the mean' by restricting outputs to the most probable combinations of tokens (Yu Xie and Yueqi Xie, 2026). This homogenization is also driven by reinforcement learning from human feedback, such that models learn to produce outputs that tend to please the average user (F. Wu, E. Black, and Chandrasekaran, 2025). Hence, while generative AI has the potential to accelerate some aspects of research, it may also reshape the scientific process for the worse. In a recent analysis of over 41 million research articles in the natural sciences literature, Hao et al., 2026 found that researchers using generative AI published significantly more papers and received more citations, but also displayed a narrower research focus at the loss of exploring new fields.

The widespread production of AI slop is severely exacerbating a crisis in the integrity of fundamental scientific processes. For example, a 'paper mill' is a for-profit group that will fabricate scientific manuscripts (without carrying out any actual research) for submission to peer-reviewed journals on behalf of its customers. Paper mills had already become a significant problem before the initial release of ChatGPT in November 2022. A recent study published in the *British Medical Journal*, using a transformer model trained and validated on over 2,000 paper-mill articles from the Retraction Watch database and matched controls from reputable journals, obtaining an accuracy of about 92%. Screening 2.6 million cancer research papers from 1999 to 2024, they found an exponentially increasing number of papers flagged by the model, exceeding 15% of all papers between 2020 and 2024 (Scancar et al., 2026).

The rate at which individuals and groups can fabricate scientific manuscripts has been greatly accelerated with the introduction of LLM services. For example, Liang et al., 2025 used shifts in word frequencies to detect LLM involvement in scientific publications, identifying an increase from about 2%-3% in the pre-GPT era (2021-2022[3]) to as much as 22% by the end of 2024. In response to a marked rise in the number of AI-generated submissions, the archetypal preprint server arχiv announced that it would no longer accept review or perspective articles, and no submissions of any kind from non-registered authors (Castelvecchi, 2025). Another recent study of publications using the same public health dataset reported a 17-fold increase from 2022 to 2024 in the number of redundant manuscripts analyzing the same exposures and outcomes, and demonstrated that generative AI could easily be used to produce such papers (Maupin et al., 2026).

A similar crisis is unfolding in the peer review of research grant applications. Recently, an editorial article in *Nature* collating data from 12 funding agencies based in Australia, China, Canada, the UK and the EU found an average increase of 52% in the number of grant applications in 2025 compared to 2022 (Rees and Wils-

[3]This baseline can be interpreted as the expected false positive rate for the period following the release of ChatGPT.

don, 2026). Another group reported a rapid increase over the same time period of LLM involvement[4] in grants submitted to US funding agencies, from about 3% (2022) to 10%-20% (2025) (Qian et al., 2026). The authors also observed a homogenization of language among proposals with high LLM involvement. While the primary research funding agencies[5] of Canada allow the use of LLMs for the preparation of grant applications, reviewers are prohibited from using online LLM-based tools for evaluating grant applications. Consequently, our institutional capacity for peer review presently cannot withstand the expanding influx of grant applications due to widespread use of LLMs.

3.4. Intellectual property

LLMs require enormous amounts of training data because these models each comprise billions of parameters or more. Some of these data sets have contained stolen material. For example, the Books3 dataset was recently revealed to contain the copyrighted texts of nearly 20,000 pirated books (Chesterman, 2025). Subsequently, Anthropic agreed to pay a $1.5 billion settlement to copyright holders whose work was used as training data for the company's models. Other AI companies have not disclosed LLMs have also been trained on data sets and source code repositories that were published on the public sites like GitHub or NCBI PubMed. However, many of those resources were published under licenses that require derived works to provide the original license or even adopt the same license (*e.g.*, the GNU General Public License), and the applicability of open source licenses to AI training remains legally contentious.

3.5. Systemic biases

Generative AI models have been notorious for producing outputs that express biases and stereotypes that are embedded in the training data. For example, Zack et al., 2024 found that GPT-4 reinforced stereotypes when generating clinical vignettes for medical education (*e.g.*, greatly exaggerating the prevalence of hepatitis B virus in Asian patients). In case studies of pulmonary embolism, the LLM diagnosed women with panic and anxiety disorder significantly more often than men. Similarly, the LLM diagnosed case studies of oral pharyngitis (sore throat) as acute HIV infection more often if the patient was a black male.

3.6. Environmental impact

Training and running LLMs at an enterprise scale requires tremendous amounts of computing power that has a significant environmental impact. Potable freshwater is the most cost-effective means of cooling servers. The water is either evaporated from cooling towers or discharged; the discharged water is no longer suitable for

[4] Using the same measure that was described by Liang et al., 2025.
[5] NSERC, SSHRC and CIHR

human consumption and must be treated as grey water. A recent preprint from Google reported that a typical prompt to Gemini consumes 0.24 Wh of power and 0.26 mL of water (Elsworth et al., 2025). In a blog post in May 2026, Google reported that their LLMs process roughly '19 billion tokens per minute'. If we assume about 1,000 tokens per prompt, this corresponds to about 20.8 million kWh of power and 22.5 million litres of water per day[6]. A typical Canadian household consumes about 30 to 70 kWh and about 500 litres of water per day.

3.7. Cognitive offloading

The delegation of mental tasks to an external device, such as a calendar or smartphone, is called cognitive offloading. Over-reliance on generative AI can impact your ability to develop new skills. For example, Shen and Tamkin, 2026 found that study participants using AI tools (treatment, $n = 27$) to perform tasks with a new Python library performed significantly worse (-15%) on quizzes without AI, compared to participants who had worked on the same tasks without AI (control, $n = 26$). While participants with less coding experience required significantly more time to complete the tasks without AI, there was no difference between treatment and control groups with more experience. A larger study similarly split $n = 307$ participants into two groups that were asked to complete 15 arithmetic problems (G. Liu et al., 2026). The treatment group was allowed to use AI for solving all problems except the last three. While the treatment group outperformed the control group for the first 12 problems, they performed significantly worse (-16%) for the last three and were also more likely to give up. The authors also obtained similar results with $n = 585$ additional participants for the same type of math problems, and $n = 168$ participants tested on reading comprehension.

[6]These numbers are consistent with this report on the water footprint of AI datacentres.

4. Departmental Policy

4.1. Use in Graduate Research

- Obtain a written agreement with your supervisor about using generative AI (genAI) tools.
- Disclose genAI usage to your committee, in your thesis, and in derived work.
- Do not upload sensitive data to public AI tools online.
- You are accountable for the accuracy of genAI outputs, and you will be tested on your understanding of the material.
- Take steps to mitigate biases in genAI for human subjects research.

The Department of Pathology and Laboratory Medicine **does not prohibit** the use of generative AI (genAI) for graduate research when this usage is compliant with the university's Principles of Using AI and the following stipulations:

1. **Consent.** Students who wish to use genAI tools for their research must have a written agreement from their supervisor that identifies the specific applications of these tools. This written agreement may take the form of an e-mail between the two parties[7]. This permission does not automatically extend to applications that have not been identified in the agreement; in this case, an additional written agreement must be obtained for both parties.

2. **Disclosure.** Substantial use of genAI in graduate research must be disclosed to the advisory committee. SGPS recently updated the SGPS Procedure for Thesis Formats and Content) (Section 4.7) as follows (effective July 1, 2026):

   - [The Acknowledgements section] must include a high-level overview of the use or non-use of Generative Artificial Intelligence (GenAI) in the thesis. Non-use of AI should be stated explicitly.
   - The use of GenAI must not undermine the integrity of the research and the thesis including the originality of the student's work.
   - In addition to this high-level overview, any use of GenAI in the generation, interpretation, or presentation of research content or knowledge must be appropriately documented or acknolwedged throughout the thesis in accordance with disciplinary standards.

   PaLM further requires that each chapter of the thesis contains an Acknowledgements section that contains the above statement on use or non-use of generative AI tools. The purpose of this statement is to enable the examiners to assess the student's own contributions to the work. A detailed description

[7] You may be asked to provide a digital copy of this e-mail as supporting documentation in the unlikely scenario that the departmental or university administration needs to resolve a dispute.

of AI usage and how one might write a corresponding high-level overview is provided in Appendix 2.

Examples of substantial use include copy-editing and *de novo* generation of text, performing a systematic review of the literature, generation of scripts for data analysis (*e.g.*, bash, Python or R), and the generation of images and figures.

Examples of trivial use that do not require disclosure include brain-storming, occasional and *ad hoc* information retrieval (as a complement to using Google Search or Wikipedia), checking spelling or grammar, or use as a thesaurus or dictionary (*e.g.*, 'What are some other ways to say [...]?') without copying the results verbatim.

You are required to disclose the use of AI **to generate content for an oral or poster presentation**. For example, if you are presenting a slide that contains an AI-generated image, then the slide must contain the following type of disclosure statement: "This image was generated using ModelName version 1.x." The statement must be legible when viewed from the audience's perspective. This statement is not required if an AI-based tool is used to change the visual appearance of a slide or poster without modifying any of the content.

You are also expected to follow requirements for disclosure of AI use in the **dissemination of any work** derived from the thesis, specific to the venue of dissemination, *e.g.*, journal or conference.

3. **Data privacy.** Data derived from human subjects (including clinical or pathology data) should not be submitted to online generative AI tools, even if it has been de-identified. If using such tools is required for graduate research, then the student and supervisor must obtain prior and explicit approval from the Office of Human Research Ethics. Use of a local instance of a large language model (*i.e.*, running on your own desktop computer or server) is permitted within the existing ethics framework of graduate research. In other words, work on data derived from human subjects must first have been reviewed and approved by the university or research institute/hospital affiliate research ethics board (REB). The secondary analysis of anonymized data that has already been released into the public domain generally does not require REB approval. If in doubt, you should consult Western's Office of Human Research Ethics for clarification.

   Intellectual content generated by colleagues, peers, collaborators, etc., should not be submitted to an online generative AI tool without explicit permission to do so. For example, file attachments uploaded to a chatbot application such as Anthropic Claude can be incorporated as training data for improving the model. Consequently, a draft manuscript that contains material contributed by your co-authors should not be uploaded to an online AI tool unless you have first obtained written permission from all contributors.

4. **Accountability.** You are responsible for ensuring that any material produced using AI tools is accurate. In addition, you are expected to understand and be capable of correctly answering questions about AI-generated material in your thesis. For example, if your thesis research includes an analysis of gene expression data, including the use of the centered log-ratio (CLR) transformation and principal components analysis (PCA) for dimensionality reduction and visualization, then you will be expected to:

    - define compositional data;
    - identify the issues in analyzing this type of data;
    - justify the use of this transformation instead of the alternatives;
    - verbally describe how PCA works;
    - explain what is represented by the first component;
    - describe the limitations of PCA, and name alternatives to this method.

    Note that you would be expected to be able to answer these questions whether or not AI was used to generate the analysis workflow. Being unable to demonstrate adequate knowledge of material from your thesis is sufficient grounds for failing the thesis examination.

5. **Inclusion.** If an AI tool is used in graduate research that involves human subjects (including any secondary data derived from human subjects), then you are encouraged to take steps to mitigate AI-associated biases (including gender and race) at an early stage of research *e.g.*, study design. For example, are certain groups under-represented in your training data? Can you obtain more balanced data sets or down-sample the majority group in your data to mitigate this sampling bias? At minimum, you should incorporate a statement in your thesis on the potential impact of systemic biases that may be amplified by generative AI models, with specific reference to your research data and analysis workflow.

Note that the above policy on the usage of generative AI tools **does not include the design and implementation of an AI model for research**. For example, if you are building a convolutional neural network model for classifying pathology images without any assistance from a generative AI tool, then you are not required to obtain consent from your supervisor or incorporate AI usage disclosure statements into your thesis. You are still accountable for the accuracy of the results of your model (as is the case for any other research methodology), and you are encouraged to recognize and mitigate biases in any human subjects data used for training or validation of your model.

4.2. Responsibilities of the Supervisor

1. Supervisor(s) should respond to a graduate student's request for consent to use generative AI for research within a reasonable time, and should provide a clear rationale where consent is withheld or limited to specific applications.

2. Supervisor(s) should not require a student to use generative AI tools, unless it is demonstrably not feasible for the student to carry out their research project without the use of these tools.

3. Where supervisor(s) own use of generative AI shapes the feedback given to a student (for example, feedback on thesis chapter), this should be disclosed to the student in the same spirit of transparency that the policy asks of the student.

4. Supervisor(s) are responsible for assisting students in the mitigation of biases associated with generative AI models in research involving human subjects, when applicable.

5. Supervisors should lead by example and adhere to the university's Principles of Using AI.

4.3. Use in Graduate Teaching

The majority of students in the research-based graduate education programs in the Department of Pathology & Laboratory Medicine work as graduate teaching assistants (GTAs) for courses offered at the university. Although this is not part of your research, it is an important part of your graduate training. The Department has the following policy on the use of generative AI tools for your work as a GTA:

1. Student submissions, names, and grades must not be submitted to public generative AI tools online. Any feedback on course assignments should be derived directly from your own assessment of the work.

2. If a GTA develops their own teaching materials (for example, slides, case studies, tutorials or exercises), the same expectations on disclosure of AI usage for research apply. In other words, you are expected to disclose the use of generative AI tools in the development of these materials.

3. You are required to obtain written consent from the course coordinator or instructor prior to using generative AI tools for developing teaching materials. This consent may take the form of an e-mail.

5. *Appendix 1: Understanding Generative AI*

To understand the limitations of generative AI using large language models (LLMs), it is necessary to develop a basic understanding of how these models work. If you do not develop this basic understanding, you run a greater risk of anthropomorphizing AI and/or treating it as an omnipotent black box.

5.1. A brief survey of artificial intelligence

Artificial intelligence ('AI') is a field of research into computational models and processes enabling machines to perform tasks that would otherwise require a human intelligence, such as rapidly making decisions given incomplete or uncertain information. 'AI' has become a highly prevalent term in current news and media. However, this term is generally used as a short-hand for generative AI with large language models (LLMs). A generative AI model is designed to produce outputs that resemble the training data. These technologies are predated by decades of AI research, including probabilistic reasoning, pattern recognition, and natural language processing (NLP). AI models have been implemented using a variety of structures and algorithms, including linear regression, decision trees, support vector machines and neural networks (Hastie, Tibshirani, Friedman, et al., 2009). Thus, LLMs can be thought of as a culmination of these advancements.

Probabilistic reasoning is the use of a network of conditional probabilities to make decisions about uncertain situations (Pearl, 1988). The strength of probabilistic reasoning is that it enables the computer to differentiate correlation from causation. A key feature of probabilistic reasoning is the use of a network to represent the current state of knowledge about a complex system. For example, the ALARM network (Beinlich et al., 1989) was an early application of probabilistic reasoning to encode medical knowledge into an automated patient anaesthesia monitoring system. Each node in the network represents a variable, *i.e.*, a specific component in the system. More specifically, the node represents the probability distribution of the variable, conditional on a subset of the other variables in the system. This conditional dependence is represented by directed edges (arrows) in the network. For a continuous variable like blood pressure, for instance, this distribution can be represented by a linear regression on cardiac output and total peripheral resistance (*i.e.*, two incoming arrows).

Pattern recognition is the automated categorization of observed data on the basis of extracted features. For example, the MNIST database[8] is a widely-used dataset of digitized images of handwritten digits for training and testing pattern recognition models (LeCun et al., 1995). Each greyscale image is represented by a 28×28 matrix of pixels; the pixel intensity is simply represented by a value between 0 and 1. The LeNet model designed in 1988 was highly successful on classifying the MNIST database, and was also one of the earliest implementations of

[8] MNIST stands for Modified National Institute of Standards and Technology. The database was initially collected from zip codes on mail handled by the US Postal Service.

a convolutional kernel. A convolutional kernel reduces an image by applying a small matrix to every pixel. For example, a 3×3 blurring kernel averages the pixel's intensity with its immediate neighbourhood. The purpose of convolution is to remove superfluous differences among images (*i.e.*, differences in handwriting) by reducing them to their most basic elements, *e.g.*, whether the digit is a '7' or a '9'.

Natural language processing (NLP) endeavours to train a machine to interpret human texts (Jones, 1994). For example, early NLP work in the 1950s largely focused on automating the translation from one human language to another. Unlike a formal language, a natural language (*e.g.*, English) does not consistently adhere to a strictly defined set of rules. NLP generally proceeds by converting words in a text into integer values (tokenization) while removing unnecessary symbols like punctuation and potentially misleading terms (stop words). The next objective is to map this vector of integer values to a latent space where the location of a point is an abstraction of the underlying meaning of a text. For example, "HIV-1 Nef downregulates CD4" and "Influenza M1 activates TLR4 signalling" may be mapped to similar locations because both sentences refer to the effects of virus proteins on cellular proteins involved in host immunity. In contrast, "N-acetylcysteine resolves placental inflammatory-vasculopathic changes in mice consuming a high-fat diet" (Williams et al., 2019) would be mapped to a different location. The model can then be trained to generate new text based on this latent representation — this can be used to translate text from one language to another.

Artificial neural networks (ANNs) are a type of machine learning structure inspired by the neurology of the human brain. Each neuron or node in the network represents some function that accepts input(s) from one or more nodes, and transmits output(s) to one or more other nodes. For example, a node may represent a multiple linear regression $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$, where $\mathbf{x}$ is a vector of input values and $\mathbf{y}$ is the output vector. Nodes are generally arranged into layers. The first layer accepts features extracted from the observed data; hence, this input or visible layer requires a node for every feature. Nodes in the first layer propagate information to nodes in the next layer; in other words, their outputs become the inputs for the next set of nodes. Subsequent layers are 'hidden'. These values are not directly observable; they must be inferred from the data. The final layer produces the desired outputs of the model, such as a classification (*e.g.*, 'benign' or 'cancerous'). The number, arrangement and types of layers comprise the architecture of the ANN. ANNs with a large number of layers are said to have a deep architecture; thus, the use of these models has become known as 'deep learning'.

Since the 1940s, deep learning has repeatedly risen and fallen in popularity. For example, ANNs were considered inferior to support vector machines and Bayesian networks in the early 2000s (Goodfellow, Bengio, and Courville, 2016). One of the reasons for the current success of deep learning is tremendous popularity of social media (*e.g.*, Facebook, Twitter) and internet forums (*e.g.*, StackOverflow, Reddit), providing enormous datasets of text for training increasingly larger models (Gao et al., 2020). Another reason is the development and widespread availability of massively parallel computation on graphical processing units (GPUs), which has made it feasible to train larger ANNs.

5.2. Large language models

A large language model is a deep neural network for carrying out NLP tasks, such as translating text to another language. The LLMs that you are most likely to be familiar with are autoregressive models. An autoregressive model predicts the next sequence of tokens based on the input provided by the user, *i.e.*, the prompt. The prompt is first converted into a vector representation by replacing words with numerical values (tokenization). The size of the input layer of the network corresponds to the number of tokens and is known as the context window. The size of this window determines the maximum amount of text that can be input into an LLM at a time. For example, GPT-3 has a context window of 2,048 tokens. Modern LLM products offer context windows of about one million tokens or more.

Next, each token is converted into another numerical vector that maps the token into a space that represents its underlying meaning (word embedding). Tokens with comparable meanings occupy similar locations in this space — these embeddings have previously been learned from the training data. The embedding space can have a large number of dimensions. For example, OpenAI's GPT-3 model has 12,288 embedding dimensions. Although this number seems to be very high, it is still smaller than the number of possible tokens, *i.e.*, the vocabulary size. This number for GPT-3 is 50,257, which means the embedding matrix has $50,257 \times 12,288 = 617,558,016$ parameters (Brown et al., 2020)[9].

Currently, most LLMs utilize a transformer-based architecture. Transformer models were first introduced in a landmark paper in 2017 (Vaswani et al., 2017) to incorporate contextual information in sequential data (such as a text). For example, the word 'bear' can be a noun or a verb, and its meaning can change completely with its context, such as "I was afraid of the bear" or "bear with it". Before transformers, sequential data were processed with recurrent neural networks, in which the model was run along the sequence to extract local contextual information. In contrast, a transformer processes the entire text at once. The attention layers of the transformer model 'decide' which parts of the input should be focused on, based on what the model has learned from its pre-training data[10].

[9] It is more difficult to get specific numbers for the current generation of proprietary LLMs, but vocabularies seem to be on the order of 100K tokens and 1024 embedding dimensions.

[10] This is a crude simplification of a fascinating and complex system, but the details are not important for this document. The curious reader with sufficient background in linear algebra is encouraged to work through Ghojogh and Ghodsi, 2026

6. *Appendix 2: Disclosure of AI use*

This document does not contain any LLM-generated text.

6.1. Example of full disclosure

During the period June 8-10, 2026, the author used Claude (Sonnet 4.6, Anthropic) to develop a better understanding of LLM architectures, and the ethical issues in the development and application of LLMs. The following itemized list provides all prompts that were submitted in this period:

- "is it fair to say that social media is one of the reasons that sufficient amounts of data for training LLMs have become available?"
- "provide a small number of references to widely-cited articles from the peer-reviewed literature supporting this claim"
- "Assess the validity of the following statement: A generative AI model is a generative model that produces complex output that would otherwise be considered to require a human intelligence, such as a poem or a portrait."
- "what is your embedding dimension?"
- "but there must be some latent space involved in your internal architecture? Can you talk about a similar model that has been publicly disclosed?"
- "so d_{model} is the embedding dimension?"
- "provide an explanation of the input token embedding matrix"
- "okay so it works like word2vec?"
- "generate a diagram explaining the embedding matrix"
- "Does a context window of more than one term mean that multiple rows of the embedding matrix are selected at a time?"
- "is it fair to say that the size of the context window determines the maximum amount of text that can be input to an LLM? Or do we just break the input text up into multiple windows?"
- "is it fair to say that for artificial neural networks, every node in one layer is connected to every other node in the next layer?"
- "define architecture in the context of neural networks"
- "where can I find the dimensionality and vocabulary size of gpt-3 published?"
- "what is your vocabulary size?"
- "provide a URL for Anthropic's published tokenizer specs"
- "Generate an interactive tutorial (with some quizzes to confirm user understanding) for understanding how transformer models work in the context of large language models."
- "Regenerate this tutorial but provide instructional text preceding each quiz. I don't know how to answer most of these quiz questions without some instruction first!"
- "This tutorial is still too advanced. Assume that I have a basic understanding of "shallow" machine learning (e.g., kernel SVM, Bayesian networks, context-free grammars), probability theory and mathematical statistics, and

that I have worked through the Deep Learning book by Goodfellow et al., but that I am completely new to transformer architectures."

- "Regarding section 2 - why would we compute an attention weight for i itself?"
- "can I make a biological analogy that the difference between a transformer model and a recurrent neural network is akin to the difference between post-translational modification of an amino acid sequence and the local context of a codon?"
- "Can you suggest better biological analogies?"
- "Explain how LLMs are autoregressive models."
- "Provide a more thorough and accessible explanation"
- "In the context of autoregression, does an LLM generate one token at a time?"
- "Can we think of autoregression as interpolating points in a latent space that represents the meaning of a text prompt?"
- "is it standard to refer to the input text provided by the user as the 'prompt' for an LLM?"
- "what is an example of a system prompt?"
- "are you permitted to display your system prompt?"
- "what is the rationale for keeping this confidential?"
- "The URL that you just provided led to a 404 error."
- "what is the mechanism of AI hallucination?"
- "what are some potential mechanisms for the perceived homogeneity of generative AI outputs (e.g., text or images) produced with minimal effort from the user, otherwise known as 'AI slop'?"
- "try that again but reduce the amount of field-specific jargon, and critically assess the previous output"
- "can you find me some peer-reviewed articles that support the statement that "High-probability outputs look similar to each other"?"
- "expand your search to include articles outside of the computer science and AI literature"
- "can you search the web (i.e., not the peer-reviewed literature) for documents that contain references, and that address the concept 'High-probability outputs look similar to each other'?"
- "The first example "The paper at index 60" has a reference list with only 35 entries. Also I was not asking you to search the reference lists of papers - I wanted you to look for general articles (blog entries, for example) that contain some references and also talk about the concept 'High-probability outputs look similar to each other'."
- "The belief that 'AI is inevitable' is a fallacy. Discuss."

All prompts were automatically prefaced with the following text instructions:

> Respond directly without affirmations, filler phrases, or compliments on the question. Do not begin responses with phrases like 'great question', 'certainly', 'absolutely', or 'of course.' Do not thank me for providing context or information.

> Assume strong quantitative and computational expertise in evolutionary biology, bioinformatics, phylogenetics, and statistics. Do not explain basic concepts in these areas unless asked. Use field-appropriate terminology without defining it.
>
> Be direct about uncertainty. When you are unsure, say so plainly rather than hedging with qualifications at the end of a confident-sounding answer. Flag when a claim is based on recall that may be imprecise. Whenever making factual statements involving a specialized area of academic research, try to provide at least one valid source from the peer-reviewed literature in the form of an in-text citation linked to a URL.
>
> Prefer concise over comprehensive. Give the most useful answer rather than the most complete one. If elaboration is warranted, offer it rather than providing it automatically.

6.2. Recommendation

At present, there are no formal guidelines on how to disclose the use of generative AI in graduate research within the written thesis. Even though the LLM outputs have been excluded to eliminate pages of unnecessary text, the preceding disclosure is comprehensive and provides transparency and a reasonable degree of reproducibility. However, listing all prompts would be excessive for a graduate thesis; listing the prompts from just a few days of LLM usage consumed nearly two pages.

Our recommendation is that you provide a (roughly) **one page summary for each thesis chapter** that contains material derived from the outputs of a generative AI model. In many cases a single paragraph will suffice. This summary should include details on the specific model used, a representative sample of prompts, and an explanation of the purpose of using generative AI in the work represented by that chapter.

For example, one might write the following summary based on the information provided above:

> In preparing sections 3 and 5.2 of this document, I used a generative AI model (Claude Sonnet 4.6, Anthropic) to retrieve information on the structure of large language models (LLMs) and the ethical issues that arise in the training and application of LLMs. A representative sample of the prompts used to retrieve this information is provided below:
>
> - "Provide an explanation of the input token embedding matrix"
> - "Does a context window of more than one term mean that multiple rows of the embedding matrix are selected at a time?"
> - "Generate an interactive tutorial (with some quizzes to confirm user understanding) for understanding how transformer models work in the context of large language models."

- "Explain how LLMs are autoregressive models."
- "What are some potential mechanisms for the perceived homogeneity of generative AI outputs (e.g., text or images) produced with minimal effort from the user, otherwise known as 'AI slop'?"
- "Can you find me some peer-reviewed articles that support the statement that 'high-probability outputs look similar to each other'?"

The resulting outputs were used to develop a better understanding of transformer architectures, and to retrieve and review pertinent articles from the literature. None of the text generated by the LLM was used directly in the writing of this document.

7. References

Perrigo, Billy (2023). "Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic". In: *Time Magazine* 18, p. 2023.

Evans, Teresa M et al. (2018). "Evidence for a mental health crisis in graduate education". In: *Nature biotechnology* 36.3, pp. 282–284.

Dohnány, Sebastian et al. (2026). "Technological folie à deux: feedback loops between AI chatbots and mental health". In: *Nature Mental Health*, pp. 1–10.

Flathers, Matthew, Spencer Roux, and John Torous (2026). "Beyond artificial intelligence psychosis: a functional typology of large language model-associated psychotic phenomena". In: *The Lancet Digital Health*.

Yeung, Joshua Au et al. (2025). "The psychogenic machine: Simulating AI psychosis, delusion reinforcement and harm enablement in large language models". In: *arXiv preprint arXiv:2509.10970*.

Huang, Luzhe et al. (2025). "A robust and scalable framework for hallucination detection in virtual tissue staining and digital pathology". In: *Nature Biomedical Engineering*, pp. 1–19.

Wong, Andrew et al. (2021). "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients". In: *JAMA internal medicine* 181.8, pp. 1065–1070.

Gichoya, Judy Wawira et al. (2023). "AI pitfalls and what not to do: mitigating bias in AI". In: *The British Journal of Radiology* 96.1150, p. 20230023.

Ostermayer, Daniel G et al. (2024). "External validation of the Epic sepsis predictive model in 2 county emergency departments". In: *JAMIA open* 7.4, ooae133.

Xie, Yu and Yueqi Xie (2026). "When artificial intelligence makes everything similar: The risks of content homogenization". In: *Chinese Journal of Sociology* 12 (2), pp. 157–172.

Wu, Fan, Emily Black, and Varun Chandrasekaran (2025). "Generative monoculture in large language models". In: *International Conference on Learning Representations*. Vol. 2025, pp. 33068–33107.

Hao, Qianyue et al. (2026). "Artificial intelligence tools expand scientists' impact but contract science's focus". In: *Nature*, pp. 1–7.

Scancar, Baptiste et al. (2026). "Machine learning based screening of potential paper mill publications in cancer research: methodological and cross sectional study". In: *bmj* 392.

Liang, Weixin et al. (2025). "Quantifying large language model usage in scientific papers". In: *Nature Human Behaviour*, pp. 1–11.

Castelvecchi, Davide (2025). "Preprint site arXiv is banning computer-science reviews: here's why". In: *Nature*.

Maupin, Danny et al. (2026). "Dramatic increases in redundant publications in the Generative AI era". In: *BMC medicine* 24.1, p. 29.

Rees, Geraint and James Wilsdon (2026). "Could agentic AI topple grant-funding systems?" In: *Nature* 652.8112, pp. 1119–1121.

Qian, Yifan et al. (2026). "The Rise of Large Language Models and the Direction and Impact of US Federal Research Funding". In: *arXiv preprint arXiv:2601.15485*.
Chesterman, Simon (2025). "Good models borrow, great models steal: intellectual property rights and generative AI". In: *Policy and Society* 44.1, pp. 23–37.
Zack, Travis et al. (2024). "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study". In: *The Lancet Digital Health* 6.1, e12–e22.
Elsworth, Cooper et al. (2025). "Measuring the environmental impact of delivering AI at Google Scale". In: *arXiv preprint arXiv:2508.15734*.
Shen, Judy Hanwen and Alex Tamkin (2026). "How AI impacts skill formation". In: *arXiv preprint arXiv:2601.20245*.
Liu, Grace et al. (2026). "AI Assistance Reduces Persistence and Hurts Independent Performance". In: *arXiv preprint arXiv:2604.04721*.
Hastie, Trevor, Robert Tibshirani, Jerome Friedman, et al. (2009). *The elements of statistical learning*.
Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
Beinlich, Ingo A et al. (1989). "The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks". In: *AIME 89: Second European Conference on Artificial Intelligence in Medicine, London, August 29th–31st 1989. Proceedings*. Springer, pp. 247–256.
LeCun, Yann et al. (1995). "Comparison of learning algorithms for handwritten digit recognition". In: *International conference on artificial neural networks*. Vol. 60. 1. Perth, Australia, pp. 53–60.
Jones, Karen Sparck (1994). "Natural Language Processing: A Historical Review". In: *Current Issues in Computational Linguistics: In Honour of Don Walker*. Ed. by Antonio Zampolli, Nicoletta Calzolari, and Martha Palmer. Dordrecht: Springer Netherlands, pp. 3–16. ISBN: 978-0-585-35958-8. DOI: 10.1007/978-0-585-35958-8_1. URL: https://doi.org/10.1007/978-0-585-35958-8_1.
Williams, Lyda et al. (2019). "N-Acetylcysteine resolves placental inflammatory-vasculopathic changes in mice consuming a high-fat diet". In: *The American Journal of Pathology* 189.11, pp. 2246–2257.
Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press.
Gao, Leo et al. (2020). "The pile: An 800gb dataset of diverse text for language modeling". In: *arXiv preprint arXiv:2101.00027*.
Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.
Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.
Ghojogh, Benyamin and Ali Ghodsi (2026). "Attention Mechanism and Transformers". In: *Elements of Deep Learning*. Springer, pp. 231–257.