Getting started with chatbots in wellbeing and mental health?



A practical guide to the responsible professional use of conversational AI

Tim Vanhove, Sylvie Bernaerts, Tom Seymoens,
George Caique Gouveia Barbosa, Tom Van Daele,
Jana Verplancke & Dieter De Witte

artevelde hogeschool
GHENT UNIVERSITY
THOMAS MORE
with the support of VLAIO

Publication details

The project team

Tim Vanhove, Tom Seymoens & Jana Verplancke - People, Society & Digitalisation, Artevelde University of Applied Sciences
Sylvie Bernaerts & Tom Van Daele - Psychology and Technology, Thomas More University of Applied Sciences
George Caique Gouveia Barbosa & Dieter De Witte - IDLab, Ghent University

Contact project lead Tim Vanhove at **Tim.Vanhove@Arteveldehs.be**

Further information: **www.arteveldehogeschool.be/ai-cares**

Suggested citation

Vanhove, T., Bernaerts, S., Seymoens, T., Barbosa, G.C.G., Van Daele, T., Verplancke, J., & De Witte, D. (2026). Getting started with chatbots in wellbeing and mental health? A practical guide to the responsible professional use of chatbots. Artevelde University of Applied Sciences: Ghent. **www.arteveldehogeschool.be/ai-cares**

Generative AI was used as a supporting tool in the preparation of this text. The technology was used to generate ideas, structure and restructure content, summarise conclusions, translate and review the language of draft versions. All generated suggestions were critically assessed by the authors, checked for accuracy and adjusted where necessary.

The authors take full responsibility for the content of this publication.

Layout: Buro Knal

© 2026 Artevelde University of Applied Sciences, Thomas More University of Applied Sciences and Ghent University, with the support of VLAIO.

Contents

Publication details p. 2

Introduction p. 4

1

Understanding

What is conversational AI?

What can AI offer wellbeing and mental health? Discover how chatbots work, where their strengths lie and what the risks are. Explore existing AI tools and develop the understanding you need to approach them critically and with curiosity.

p. 7

2

Preparing

Make organisational choices and establish policy frameworks

Using AI starts not with technology, but with the choices your organisation makes. How do you decide what AI can and cannot be used for, which data can be shared, and who is responsible? Build a clear AI framework, step by step, with the support of your entire organisation.

p. 32

3

Taking action

Test, implement and improve systematically

From idea to practice: learn how to write effective prompts, set up your own chatbot, use reliable information sources and test your solution safely. Follow clear, practical steps to create an AI tool that meets your organisation's needs.

p. 49

References p. 66

Further reading p. 68



Introduction

AI chatbots have quickly become part of everyday life. They hold conversations that sound surprisingly human. This also creates new opportunities for wellbeing and mental health. At the same time, it raises fundamental questions: when is a chatbot helpful? What information is safe to share? Who is responsible for ensuring its quality and safety?

This practical guide invites organisations to move beyond simply embracing or rejecting artificial intelligence (AI) and to make informed choices about its use. You will discover what conversational AI can and cannot do, how to make clear decisions within your organisation, and how to safely test, develop and improve applications. The aim is not to build a chatbot for every possible use, but to explore where AI can add value, with the quality of professional practice as the guiding principle.

This practical guide was developed as part of the TETRA project **AI-CARES**, with the support of Flanders Innovation & Entrepreneurship (VLAIO). AI-CARES is led by Artevelde University of Applied Sciences in collaboration with Thomas More University of Applied Sciences and Ghent University. Through a learning network, we identified the needs and questions from wellbeing and mental health organisations, while also exploring opportunities highlighted by software and AI developers. Through practical use cases, we explored and tested how conversational AI could be used in these settings.

As part of the use cases, **chatbot prototypes** were trained using specialist knowledge sources from wellbeing and mental health organisations. Together with these organisations, we tested different ways of using this knowledge to improve how chatbots respond to questions about psychosocial issues. Working with real data helped us learn what already works, what remains challenging and what is needed to use AI responsibly in wellbeing and mental health services.

Six chatbot prototypes were tested:

1. With the **Information Chatbot**, we explored how a chatbot could tailor information more closely to the user's situation without compromising the quality of the content.
2. The **Triple Chatbot** showed the effects that ethical boundaries can have when discussing sensitive topics.
3. The **Training Chatbot** enabled users to practise conversations in a simulated online chat counselling session through role-play with a chatbot, while receiving tailored suggestions and feedback.
4. The **Finetuned Chatbot** showed the effects of adapting the underlying language model through further training on thousands of real conversations.
5. The **Intake Chatbot** demonstrated how information can be gathered through a chat conversation while someone is waiting for human support, so that the subsequent conversation with a practitioner can start with an overview of the relevant information.
6. With the **Matching Chatbot**, we explored how a conversation with a chatbot could connect a person's support needs with the services offered by an organisation.

Visit **https://aicares.ilabt.imec.be/** and try them yourself.

You will find more information about these chatbots later in this practical guide.

The starting point of this practical guide is simple: using chatbots is not an end or goal in itself. Chatbots can only be used responsibly when you are clear about the problem you want to solve and how a chatbot could solve it. This practical guide is not a technical or legal manual, but a resource that provides information, inspiration and guidance.

The guide has a modular structure, so you can consult its different parts separately. Start with the part that best matches your organisation's current question:

1. Do you want to understand what AI is and what opportunities and risks it brings? Start with **Part 1: Understanding**.
2. Do you want to prepare and agree on what is and is not allowed when using AI? Start with **Part 2: Preparing**.
3. Do you want to start working with your own chatbot? Start with **Part 3: Taking action**.

You will find practical checklists throughout the text. They are intended to support conversations about AI and help you make informed decisions. Do not treat them as simple tick-box exercises, but as starting points for reflection within your team. When using the checklists, involve your staff with different perspectives: practitioners, managers, ICT, communications, quality assurance and policy staff, legal advisers (a data protection officer or DPO) and, where possible, your users or members of the public.

Using AI is not a one-off decision. Technology, regulation, experience and public expectations change quickly. What is not responsible or feasible today may become possible later. At the same time, an AI application that works well today may need to be adapted as circumstances change. Use this practical guide as part of an ongoing process of learning, evaluation and improvement.



1

Understanding

What is conversational AI?

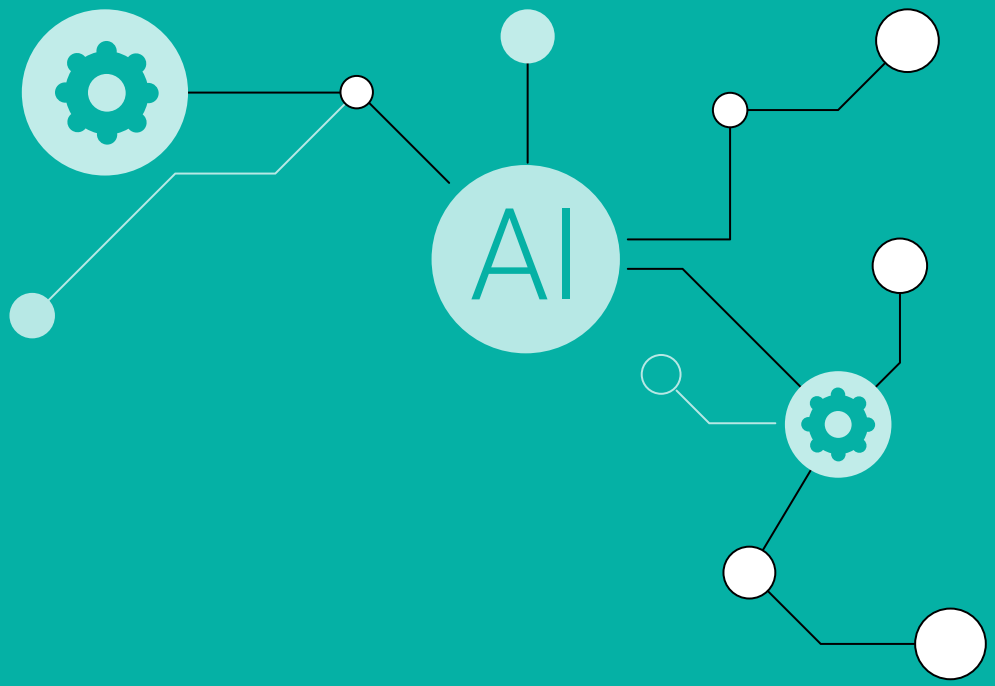
- What is AI?
- Opportunities for chatbots in wellbeing and mental health
- The risks of AI
- What does the EU AI Act mean for wellbeing and mental health?
- How can you train a language model yourself?
- Are you and your team ready for AI?
- Summary

Reading guide

What is conversational AI, and what are the key concepts? This part provides a shared language for discussing conversational AI. It is not a technical course, but an accessible explanation of what conversational AI does, where the risks lie and what opportunities it offers for wellbeing and mental health.

Read this part especially when you notice that the topic of AI raises many questions, resistance or uncertainty within your organisation. It helps you move beyond slogans such as “AI is dangerous” or “AI will take over our work”. It gives you the language to discuss what AI can and cannot do in a more nuanced way.

After reading this part, your team will be better able to understand why AI is not an ordinary search engine, why AI output must always be checked, why the persuasive nature of chatbots can be risky and why human responsibility remains central.



What is AI?

What exactly is conversational AI and how can you use it in your organisation? A restaurant analogy can help explain this. Think of conversational AI as the entire restaurant kitchen and its service. The language model is the chef and the chatbot is the waiter who talks to the customer. The customer - that is you - sees only the waiter, or the chatbot. Behind the scenes, however, there is a whole operation driven by the chef’s expertise: the language model. A chatbot is therefore the visible user interface of a conversational AI system powered by a language model. We explain the key terms more below.

Language models

Large language models are good at predicting language. They are trained on enormous amounts of text, enabling them to recognise patterns very effectively. This works so well that they can also generate language themselves, based on those patterns. Chatbots use these language models to hold context-sensitive conversations that feel very human (Stryker, 2021).

Language models are based on almost all the texts in all languages across the internet. They are therefore large, hence the name large language model (LLM). They have been trained on many billions of pieces of text and have gradually identified patterns in language. Some of these patterns are familiar to us, such as spelling and grammar rules. Others go much deeper: a language model is good at estimating which words belong together in a particular context.

Sometimes, however, a language model gives an answer that is wrong but still sounds credible. This is called hallucination. It happens because the language model does not base its response on knowledge sources, but constructs sentences using probabilities. AI-generated text can look polished and convincing, but that does not mean the content is accurate or reliable. As users, we must always think critically and distinguish between correct and incorrect information, and between reliable and unreliable sources. **A language model does not understand what it says. It simulates understanding using probabilities.**

Conversational AI and chatbots

Conversational AI is the technology that makes conversations possible with chatbots, but also with voice assistants, virtual assistants or smart speakers, such as Google Home Assistant or Alexa. A chatbot is therefore one way of using conversational AI: it is an interface through which you hold a conversation in text.

How does it work in practice? You give a chatbot input by writing words or sentences, or a “prompt”. That language is then used as inspiration or input for the chatbot to create new output. Each time, the model has to estimate the most logical next piece of text based on statistical probability. Something new is always created or generated by this next-word prediction: the text is always entirely new, rather than copied and pasted from the internet.

Why does a conversation with a chatbot such as ChatGPT sometimes feel like talking to a person? A chatbot does not understand language in the way people do, but it is exceptionally good at predicting words. This allows it to formulate answers that sound natural, fluent and convincing.

GPT: Generative Pre-trained Transformers

What does the GPT in ChatGPT actually stand for?

- **Generative:** the language model generates new text each time.
- **Pre-trained:** the language model has already been trained on a very large amount of text.
- **Transformer:** the language model compares words with one another to understand the context.

The use of transformers is a key technique in a language model. A transformer allows a language model to do more than look at words one by one: it tries to understand, simultaneously, how all the words might be connected. It identifies relationships between words and can therefore capture context much more effectively. Just as people in a conversation unconsciously give more “attention” to certain words to construct meaning, transformers do this mathematically. They weigh each word against every other word in a sentence or paragraph to determine which words are important for the meaning of that particular word. This allows them to take context into account when determining a word’s meaning.

As a result, a language model can establish complex relationships, such as who says what, what matters in a sentence and which sequence makes sense. It looks at all the words in a sentence at the same time to understand how they relate to one another and also draws on the context of earlier sentences.

For example: “The pizza came out of the oven and it tasted delicious.” As a person, you immediately and naturally understand that it is the pizza, not the oven, that tastes delicious. A transformer searches its patterns to see how often “oven” and “delicious” occur together, and how often “pizza” and “delicious” do.

Based on the probabilities in those patterns, it quickly chooses which words fit together best. In this way, it understands the necessary context: pizza can taste good, an oven less so.

Without transformers, AI models struggle with language. Before transformers existed, AI systems found it difficult to follow longer conversations. They forgot what you had said earlier or gave strange, incoherent answers. Thanks to transformers, language models can maintain the thread of a conversation, remember context and generate sentences that are both grammatically correct and coherent in content. The AI model therefore recognises how words fit into the bigger picture.

The first therapeutic “chatbot” already existed in 1966

The origins of conversational AI date back to the 1960s, when experiments were already taking place with chatbots that could formulate answers automatically. The best-known example is the psychotherapeutic ‘chatbot’ **ELIZA**, described by researcher Joseph Weizenbaum (1966). In response to “I feel sad”, for example, the chatbot could reply: “Why do you feel sad?” This seems simple today, but it was revolutionary at the time. ELIZA did not use a language model, but thousands of pre-programmed rules. The chatbot recognised certain words or sentence structures and linked them to a response. This sometimes works, but not always. ELIZA is still available: try it and decide for yourself whether you would still call this AI today.

https://www.masswerk.at/elizabot/

Opportunities for chatbots in wellbeing and mental health

Administrative workloads and staff shortages, rising demand for support and long waiting times in wellbeing and mental health services are stimulating the search for new forms of support. Conversational AI is increasingly presented as a promising technology because its interactions appear empathic and human. AI technology is evolving very quickly, and quite a few applications are already available on the market. However, scientific evidence of their benefits or effects is often still lacking.

Here, we provide an overview of the possibilities for AI in wellbeing and mental health. We show how it can support and complement human action, rather than replace it.

1. Potential AI tools for professionals

AI can support wellbeing and mental health professionals in their daily tasks, particularly as a tool for **language, structure and thinking**. It can help turn difficult texts into plain language, reduce jargon, rewrite texts at B1 level using everyday words, short sentences and an active writing style, summarise long documents, or make emails clearer or more formal. It remains important that the professional writes the initial draft, carries out the final check and does not enter personal data or confidential information.

AI can act as a sounding board for **thinking and brainstorming**. Professionals can use it to gather ideas for activities, creative titles, methods or possible approaches. AI provides inspiration, but the professional decides what is feasible, appropriate and responsible in the specific context.

Translation tools such as DeepL or Google Translate can also help translate leaflets, texts or conversations. AI allows them to take greater account of context and meaning. Errors remain possible, however, because the correct meaning in a particular context is not always clear. Use these tools mainly when you can check the target language to some extent. Also be transparent with the client about your use of a translation tool.

1. Potential AI tools for professionals

AI tools for administrative support

AI tools for practitioner training and knowledge support

AI tools to support triage and planning

AI tools to support professional practice

Many practitioners spend a considerable amount of time on administration, although they would rather spend that time with clients or users. Conversational AI could reduce this burden. AI can act as a **text editor**: turning rough notes into a report, summarising, rewriting, structuring or shortening texts, identifying action points, allocating tasks or bringing information together under predefined headings. It remains important that staff think for themselves, determine the structure and provide the initial substantive input. AI then builds on professional thinking rather than taking it over.

Specialised AI tools can also help to **reduce administrative workloads and support everyday professional practice**. Based on carefully anonymised notes, AI can help with reporting and case preparation. It can summarise a conversation, organise themes, draft a report or produce a task list for those involved. It can also suggest additions to an approach that the practitioner has already developed.

The key principle is that AI supports professional judgement rather than replacing it. Professionals should first make their own assessment, provide relevant context, critically review AI-generated content and make any necessary changes. Cases involving safety concerns, risks to wellbeing or ethical dilemmas still require careful discussion with colleagues and supervisors. Speech-to-text tools can record and transcribe conversations and turn them into reports tailored to the way an organisation works. Such applications can save time, but require clear agreements on consent, privacy, professional duty of confidentiality, storage and checking the report.

The chatbot **Maya** is an example of a system integrated into electronic patient records. It supports healthcare professionals in finding information, navigating systems and carrying out administrative tasks such as scheduling appointments. For example, the application means nurses no longer have to telephone patients after surgery - calls that often went unanswered and were therefore time-consuming - because patients can provide the necessary information through the app.

1. Potential AI tools for professionals

AI tools for administrative support

AI tools for practitioner training and knowledge support

AI tools to support triage and planning

AI tools to support professional practice

A chatbot can take on the role of a client in a **practice conversation** with staff in training. Practising communication skills is often time-consuming, and these role-plays can sometimes feel uncomfortable. An instructed AI-trainer can help practitioners develop both appropriate communication techniques and confidence in their conversational skills. Afterwards, the tool can provide feedback on the conversation, for example on tone, structure, questioning or responsiveness to the client. These digital training assistants are particularly useful because they allow staff to practise realistic situations safely, without placing pressure on real clients or trainers. Human supervision is still needed: AI feedback can support learning but does not replace professional assessment or peer reflection. In the AI-CARES project, we experimented with a training tool for the social work umbrella organisation SAM and several chat counselling services that can help both new and experienced practitioners practise their chat communication skills.

Chatbots can also act as **information assistants**, helping staff find information in frequently used sources, such as referral contacts, legal information or information about social services.



1. Potential AI tools for professionals

AI tools for administrative support

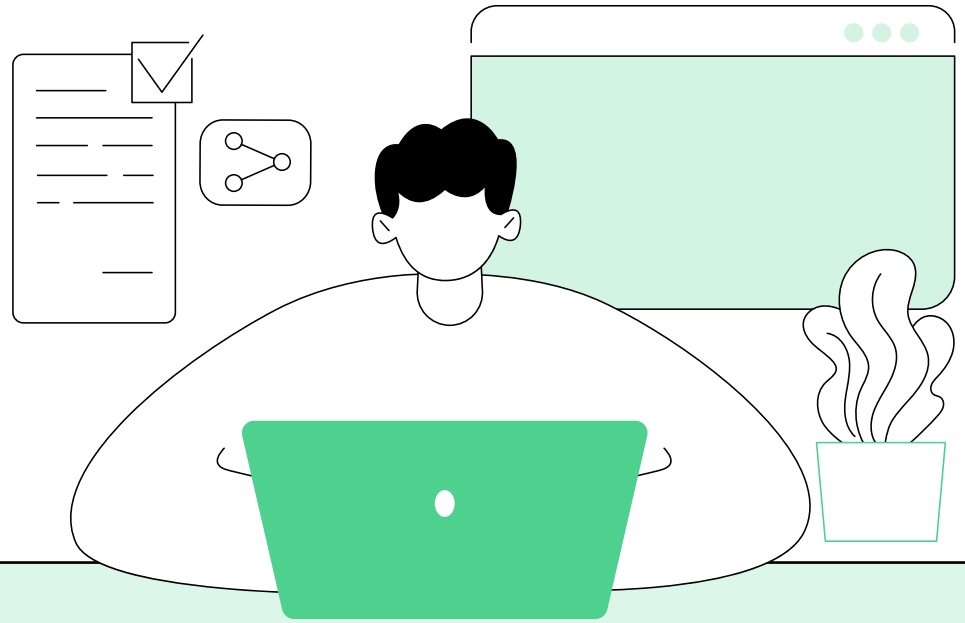
AI tools for practitioner training and knowledge support

AI tools to support triage and planning

AI tools to support professional practice

Conversational AI can **automate** repetitive tasks and potentially reduce cognitive workload. A practical example in mental health care is the UK's **Limbic Access**, which supports intake processes and clinical assessment. Limbic Access uses an AI chatbot that can be embedded in an organisation's website. It gathers information and assists with triage, helping professionals work more efficiently and arrive at an initial clinical assessment more quickly.

Another example is **Inquira Health**. This platform uses an AI voice assistant as a virtual receptionist that can answer calls, process cancellations, reduce missed appointments and support intake, allowing staff to focus on clients who are physically present.



1. Potential AI tools for professionals

AI tools for administrative support

AI tools for practitioner training and knowledge support

AI tools to support triage and planning

AI tools to support professional practice



AI can also provide support in professional practices, provided the technology is used for appropriate tasks. These should focus on support without taking over the practitioner's professional responsibility.

Chatbots can act as **sounding boards** for preparation and reflection. AI can help prepare for a difficult conversation, adapt psychoeducation to the target group, draft an initial conversation guide or explore possible points to consider. The practitioner remains the person who judges what is appropriate and safe. AI acts as a sounding board: it helps find suitable language, explore options and provide structure, but does not make decisions.

The Belgian **Psy-AID project**, run by Artevelde University of Applied Sciences and Howest, investigates how AI-generated images can make psychoeducation more closely tailored to clients' needs. This makes psychoeducational information more accessible and inclusive.

Although these applications require further research, chatbots are also being explored as tools to support clients between sessions with a professional. Michael Heinz and colleagues at Dartmouth College, for example, demonstrated how the chatbot **Therabot** could effectively improve symptoms of depression, eating disorders and anxiety (Heinz et al., 2025). An important qualification is that these applications must have the necessary **medical device certification** to be used in Europe.

2. Potential AI tools for the public

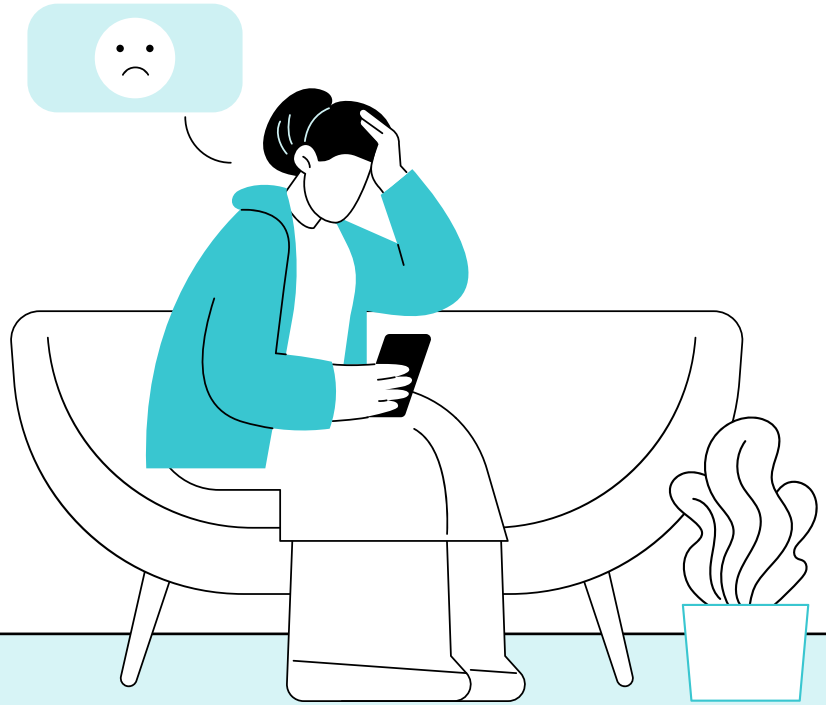
AI information assistants

AI tools to help people manage their wellbeing and health

Wellbeing and mental health services are often experienced as complex and difficult to access. Conversational AI can act as a guide, helping people find their way to the right support.

The chatbot **Agathe/Agatha**, developed for the OCMW of Sint-Agatha-Berchem, a local public social welfare centre in Belgium, supports residents with administrative and social questions and helps them navigate their entitlements and available services. In a similar initiative, Brussels University Hospital launched the **virtual assistant Aurore**, mainly to bridge language barriers and make information more accessible to people who speak other languages.

In the AI-CARES project, De Ambrassade explored using an AI information assistant instead of a traditional search bar to make information from its youth website **WatWat** more accessible and better suited to young people.



2. Potential AI tools for the public

AI information assistants

AI tools to help people manage their wellbeing and health

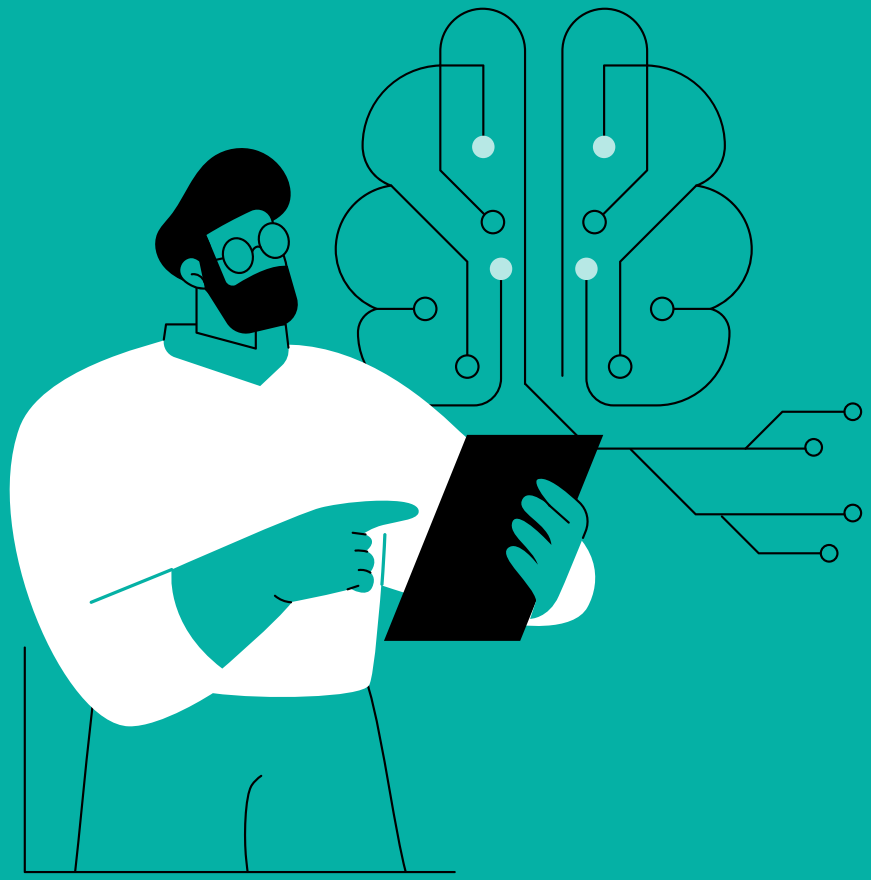
Although people already use general-purpose AI chatbots such as ChatGPT and Copilot to ask wellbeing-related questions and look up medical information, these chatbots have not been appropriately trained for these purposes. In response to this need, some companies focus specifically on medical support for the public. **Symptom checkers** such as **Ada Health** allow users to analyse their symptoms and obtain an initial indication of possible causes and next steps. This could potentially support patients, help direct requests for care more appropriately and avoid unnecessary consultations. Some applications go a step further. Cerina Companion, with **the chatbot Hedgie**, aims to support patients on waiting lists through cognitive behavioural therapy, symptom tracking and check-ins. However, despite collaboration with healthcare professionals and training using clinical data, the safety and effectiveness of these applications have often not been sufficiently studied.

For blind and **partially sighted people**, apps such as EMVI and Be My Eyes provide support by using AI to describe their surroundings or connecting users with volunteers or AI through video.

Several organisations providing an **online chat counselling service** have already explored how chatbots could improve their services. The Dutch suicide prevention service 113.nl, for example, experimented with AI support for its helpline, including analysing chat conversations and supporting and training practitioners. In Finland, **Sekasin Chat** offers young people an accessible chat service, where AI can help during busy periods or when the service is unavailable. The chatbot can acknowledge a person's experience, provide basic information and refer them to support, but does not replace the human practitioner. When there are concerning signs of suicidality, the chatbot must recognise the limits of its role and refer the user to human support. Alongside the Sekasin chatbot for children and young people, Otherkind developed a **Moodbot** for adults with questions about wellbeing and emotions.

Chatbots are often free, available at all times, do not require travel and offer a sense of anonymity. This lowers barriers, particularly on sensitive topics such as mental health. Applications such as **Ebb (Headspace)** and **Ello (HelloBetter)** are marketed as AI companions for mental wellbeing, often as a way of avoiding official classification as a medical device and therefore the associated legislation. These apps offer users structured conversations, psychoeducation and exercises based on meditation and breathing techniques.

The risks of AI



Although there are several interesting AI tools for wellbeing and mental health, it is important to understand their limitations and risks. In these fields in particular, errors or misuse can have a major impact on users, professionals and organisations. The central point is that conversational AI can be powerful and useful, but in wellbeing and mental health it must always operate within clear professional, ethical and organisational boundaries.

This section discusses the main risks identified in research. Risks and limitations also emerged in the AI-CARES learning network and use cases. Although researchers are now conducting many studies on the usability, safety and effectiveness of chatbots, most conversational AI applications have not yet been validated scientifically in the real world. The risks described below are therefore those demonstrated at the time of writing (September 2026). Developments in AI and new versions of language models follow one another faster than research can keep up. New language models and AI applications may address these risks or introduce new ones.

Always discuss the use of conversational AI within your organisation with experts such as your data protection officer (DPO) or legal adviser.

1

Lack of transparency

For many AI systems, it is unclear which data the model was trained on, which instructions it follows, what quality controls and limitations are built in, and which commercial interests are involved. This makes it difficult to assess the assumptions and limitations behind an AI-generated answer. This is problematic for wellbeing and mental health, because professionals and users need to be able to assess what information, recommendations or advice are based on.

There may also be uncertainty about how much AI was used to create texts, images, presentations or other content. What was created by a person, and what was produced with the help of a language model? This requires transparency.

HOW TO ADDRESS THIS

Inform yourself thoroughly

Find out about the training data and guardrails of AI tools. At present, this is often not possible for an individual or organisation, but it may become possible as models evolve.

When your organisation develops its own chatbot, make information publicly available about the technology used, the instructions or system prompts that apply, and how quality is monitored.

Decide how your organisation should disclose its own use of AI. The greater AI's contribution to the content and the greater the potential impact on users, the more important transparency becomes. Establish an organisational guideline. One possible standard statement is: "Some of this content was written or edited using generative AI. The content was checked by a member of staff and approved or adjusted where necessary."



HOW TO ADDRESS THIS

2

Lack of reliability

AI recognises patterns but does not understand them. Generative AI remains susceptible to hallucinations: answers that sound convincing but are factually wrong, incomplete or invented. Biases in the training data may also appear in the output. This is particularly sensitive in wellbeing and mental health, where accuracy and safety are crucial.

Check the output.

Stay in control and take responsibility by thoroughly reviewing all output for accuracy and completeness and reworking it using your own expertise, experience and reliable knowledge sources. Check for implicit bias as well. In other words, rewrite the text and make it your own. Your organisation remains responsible for the thinking.

HOW TO ADDRESS THIS

3

Too little or too much human-like behaviour

AI can convincingly imitate empathy, connection and understanding, but it feels nothing and does not understand vulnerability in the way a person can. This creates a double risk. Too little empathy can feel cold or unsafe. Too much or inappropriate empathy can make people feel they are receiving genuine help, while warning signs are missed.
In situations involving crisis, suicidality, violence, psychological distress or concerns about safety or wellbeing, this can mean that problem exploration, crisis assessment or referral is missing.

Protect the human relationship

Using general-purpose chatbots in situations involving crisis, suicidal thoughts, violence, psychological distress, or concerns about safety or wellbeing, is not without danger for users. Legally, responsibility for this lies with the users themselves, because they choose to communicate on these topics.

When your organisation develops its own specialist chatbot on mental health, assess whether AI can respond to sensitive topics in a way that is appropriate, safe and professionally responsible, because the organisation is legally responsible when the chatbot is specifically aimed at mental health. Use clear safety instructions and appropriate human oversight. Make it clear that users are communicating with AI.

4

GDPR, privacy and professional duty of confidentiality

Confidential information or information that affects people's privacy must never be entered into general-purpose AI tools. Names, addresses, dates of birth, medical or legal information, organisation names or recognisable case details can make individuals identifiable, and this is not permitted under the General Data Protection Regulation (GDPR) of the European Union.

HOW TO ADDRESS THIS

Provide guidelines organise and raise awareness

Provide organisational guidelines for professional AI use and make all staff aware of the dangers of sharing sensitive data. Prohibit the uploading of confidential, personal or privacy-sensitive information.

HOW TO ADDRESS THIS

Clarify professional AI use

Set up clear organisational guidelines: what is and is not allowed, under which conditions, with which tools, what checks are needed and how transparency should be ensured.

5

Shadow AI use without organisational guidelines

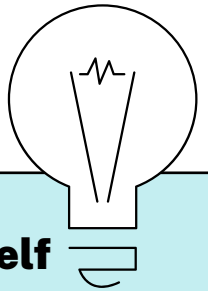
Staff may use AI professionally without a shared framework. They then decide for themselves which tool to use, what information to enter and how to assess the output. This creates risks for privacy, quality and ethics.

6

Copyright infringement

AI tools are trained on large amounts of text. It is not always clear whether all the sources used were freely available or protected by copyright. AI output may also unintentionally resemble existing protected material.

HOW TO ADDRESS THIS



Read the sources yourself

Check the cited sources and whether they may be freely reproduced, acknowledge those sources and agree on who owns material created with AI.

7

Loss of professional responsibility

AI may support professionals but must not take over their professional autonomy or ultimate responsibility. AI provides support, but it does not make decisions, however convincing it may sound. When staff use AI too often or too automatically, their own professional judgement can weaken. AI can take over their thinking. This is risky in wellbeing and mental health, because quality depends not only on accurate information, but also on professional sensitivity, contextual knowledge and ethical judgement.

HOW TO ADDRESS THIS

People must always remain in control

AI is only a tool. It must never make diagnoses, formulate hypotheses about clients or propose decisions that affect the rights, wellbeing or health of those involved. AI must not replace reflection, consultation, supervision or professional expertise.

8

Inadequate security and data sovereignty

Particularly in sensitive fields such as wellbeing and mental health, it is important to establish where data is stored and processed, who can access it, which jurisdiction they fall under and what control the organisation retains over its data. If the provider is a commercial company, questions must be asked about its business model: does it profit from using the target group's data?

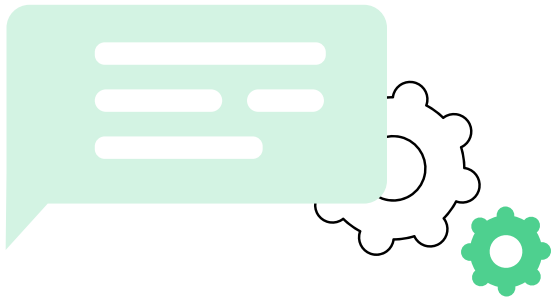
HOW TO ADDRESS THIS

Choose European safeguards where possible

Organisations should work in secure environments, preferably on EU servers and using European technology, with clear safeguards for privacy, data security and data use.

These risks do not mean that AI has no place in wellbeing and mental health. As with any new technology, the aim is not to eliminate every risk, but to recognise risks and address them consciously.

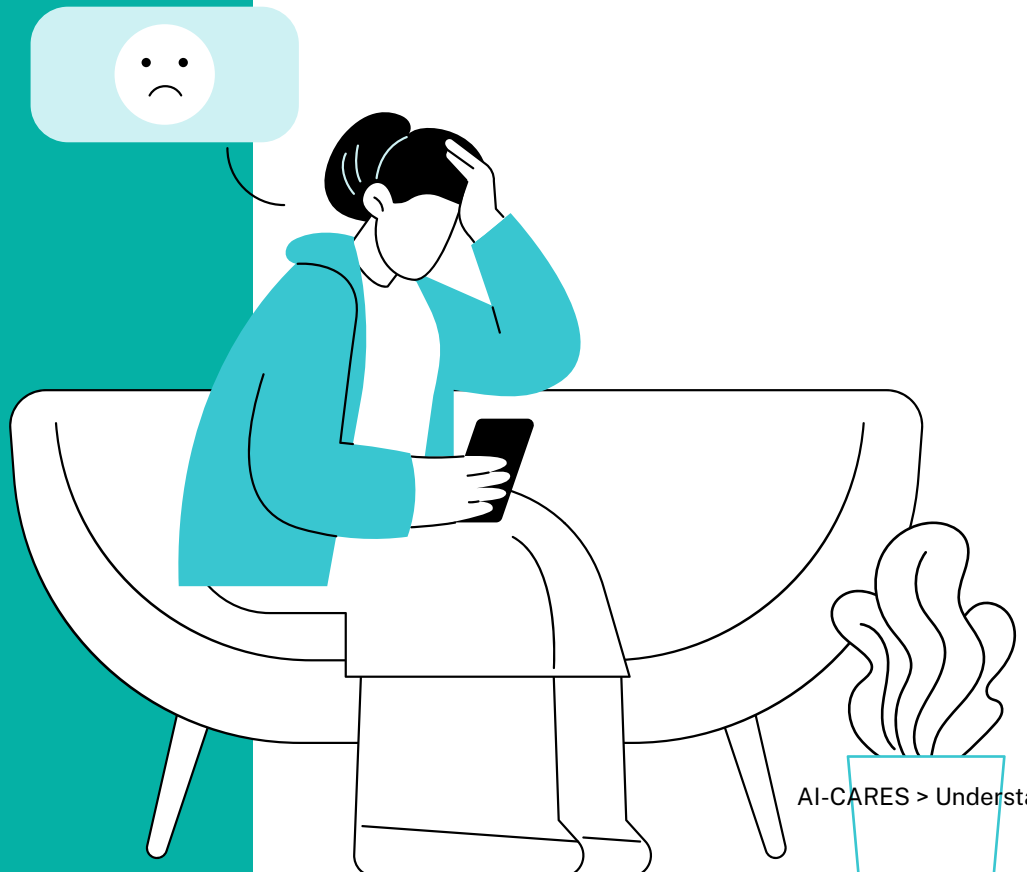
Although it may not be a direct risk to an organisation, we must also pay attention to the environmental impact as a society. Training AI models requires substantial electricity and cooling water. The impact of a single prompt may seem small, but widespread daily use adds up to a considerable environmental burden. This makes thoughtful prompting part of professional AI use: think ahead, ask clear questions, avoid excessive re-prompting and use AI only when it genuinely adds value. Take the final step yourself: do not keep prompting endlessly to perfect every detail. Accept that AI output does not have to be perfect and finish the result yourself. This helps prevent AI use from becoming automatic and keeps you thinking critically and making your own decisions.

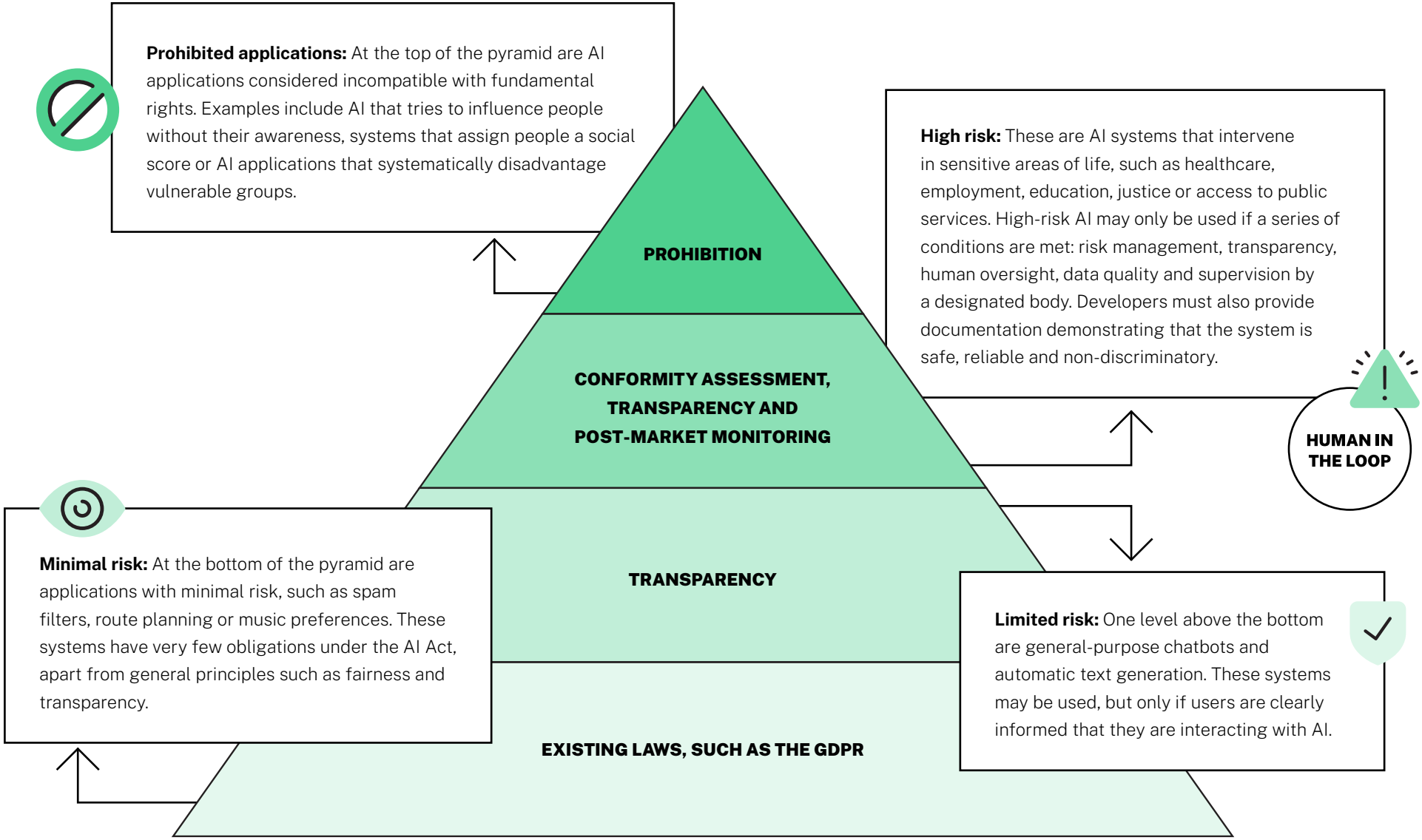


What does the EU AI Act mean for wellbeing and mental health?

The tools presented in the previous section show how AI can support wellbeing and mental health services in many ways. At the same time, these opportunities raise important questions. When may AI be used? What responsibility does an organisation bear? Which applications are acceptable and which are not? What role remains for people when AI is used? These questions are not only ethical but also legal. This is why the European Union developed a legal framework for AI. The EU AI Act sets out the rules that apply to different types of AI application.

The EU AI Act is the European Union's first comprehensive law on artificial intelligence. It takes a risk-based approach to classifying and regulating AI systems. European legislation categorises AI systems according to the risks they may pose to people.





The EU AI Act risk pyramid

A key principle of the EU AI Act is that risk is determined by an AI system's intended application, rather than its actual use. Even if some users effectively use ChatGPT as a health adviser or therapist, OpenAI's general intended purpose remains the basis for the risk assessment.

The pyramid is a legally binding instrument that determines what is allowed, what is regulated and what is prohibited. Wellbeing and mental health organisations need to identify in advance where their future AI application sits in this pyramid, because this affects design, oversight and liability. Do not make this assessment on your own: seek advice from your data protection officer (DPO) or legal experts.

One of the EU AI Act's most important principles is the requirement for **human oversight** when AI can have a major impact on people's lives. The degree of oversight can vary. In some cases, a person must explicitly approve a decision before it is carried out: human-in-the-loop. A person must then retain meaningful control over decisions proposed by an AI system. This is not passive oversight after the event, but active involvement at decision points that affect human rights, safety or wellbeing. In other cases, a person must be able to correct a decision afterwards: human-on-the-loop or human-in-command.

In wellbeing and mental health settings, this means that a chatbot specialising in psychosocial topics must always be linked to a proactive mechanism for human oversight: human-in-the-loop. Such a chatbot may not operate autonomously. Human involvement is not optional, but fundamental to the AI system functioning in an ethically and legally responsible way. For sensitive situations involving suicidal thoughts, violence or other safety concerns, simply intervening after something goes wrong is not enough. Chatbots should have clear safety boundaries, be tested for their ability to identify warning signs and provide timely access to appropriate human support. Because AI cannot reliably detect every risk, clear procedures for human intervention are essential.

Wellbeing and mental health organisations must therefore establish clear agreements on human oversight when they provide their own AI systems: who supervises, with what authority and with what level of AI literacy? This requires not only sufficient training, but also clear procedures. Examples include risk-flagging systems for sensitive topics, stop buttons, reporting channels, warnings, explanations and opportunities for escalation to a person. All of these are still at an early stage.

Make sure you involve your legal team and inform yourself thoroughly. Further information can be found on this website of the European Union: **https://ai-act-service-desk.ec.europa.eu/en**

How can you train a language model yourself?



We now have a better understanding of what a chatbot is and what opportunities and risks it presents. Those opportunities can be strengthened and the risks reduced by training a chatbot more specifically for specialist tasks. There are several ways to train or customise a chatbot. Here, we discuss three: system prompting, retrieval-augmented generation (RAG) and fine-tuning. In Part 3, we also put them to the test.

You can give a general-purpose chatbot instructions by asking a clear question or describing what you want. This is called **prompting**. For example: “Write an explanation of stress for a 14-year-old.” Everything you write is input for the chatbot. The prompt sets the language model to work, generating a plausible answer based on probabilities. The clearer and more relevant your instructions and context, the greater the chance of a useful answer. Your input therefore acts as a set of instructions or signposts for the chatbot. This is not training in itself, but it is part of using AI effectively.

Alongside the basic way of everyday prompting, there is **system prompting**. This operates more at the level of the chatbot itself. The chatbot receives general instructions about its role, tone, approach and boundaries. Through system prompting, you can specify that it should respond in a friendly and clear way, not make diagnoses, always refer users to human support when signs of crisis appear, or follow a step-by-step process. These instructions shape the chatbot’s overall behaviour. You can simply write them in words, without programming languages or complex systems.

Use few-shot prompting as well, because it is useful to give the language model a few good examples - “a few shots” - to show the kind of answer you want. This teaches the chatbot not only what to do, but how it should sound. Few-shot prompting works well and is efficient. With zero-shot prompting, you provide instructions but no examples. One-shot prompting adds one example. Many-shot prompting adds a very large number of examples.

The second technique is **RAG**, which stands for retrieval-augmented generation. In simple terms, this means generating new text using existing information or knowledge that first needs to be retrieved. RAG gives a language model access to specific information sources of your choosing, such as documents, websites, internal databases, protocols or manuals. When you ask a question, the system first retrieves relevant pieces of information from these sources and then uses that information to formulate an answer. The language model and the external knowledge sources therefore work together but remain separate: the model provides the wording, while the sources provide the information on which the answer is based.

For example, a practitioner at a CAW (a Flemish wellbeing organisation) wants to know how to respond to suspected child abuse situation but does not know the internal protocol by heart. They could look this up in the manual themselves. With a RAG-based system, however, a chatbot can help them find and summarise the relevant passages more quickly. The organisation adds documents to its internal knowledge chatbot and provides instructions through system prompting. The chatbot then searches the guidelines and formulates an answer based on their content. It does not replace professional judgement or consultation. Its main role is to help locate and organise information.

A third technique is more complex. **Fine-tuning** is the process of further training an existing language model using many additional examples or texts. The language model itself is therefore slightly modified. In effect, you teach it additional patterns so it can respond in a more appropriate style. Fine-tuning is particularly useful when you want a model to respond consistently in a particular way, for example in an organisation's tone, following a particular conversational approach or based on a large number of examples of good support conversations. RAG is more suitable for (changing) factual information, such as procedures, opening hours, legislation or internal agreements.

You can understand the **difference** between RAG and fine-tuning by comparing two types of exams. With fine-tuning, you give the model additional lessons beforehand and then ask it to answer without consulting documents: a closed-book exam. The model has been adjusted in advance using examples and draws on those learned patterns in its answers. With RAG, the model does not have to learn everything beforehand. As in an open-book exam, it can look up information in external sources, such as a manual, website or internal knowledge base. The system first retrieves relevant pieces of information from the source and uses that context to formulate an answer. Fine-tuning mainly changes the language model's own behaviour. RAG gives the model access to specific knowledge sources without changing the language model itself.

In Part 3, we compare these techniques by developing prototypes of trained chatbots.

Are you and your team ready for AI?

Wellbeing and mental health organisations have a lot of experience-based expertise. This knowledge is essential for making informed decisions about AI and developing an organisational AI policy. Technology is only a tool. Staff therefore do not all need to become AI specialists, but they do need to understand what AI can and cannot do. Only then can they critically assess its possibilities and limitations. The following checklist will help you assess how prepared your organisation is to use AI responsibly.



CHECKLIST: Do we understand AI well enough to take the next step?

Do we understand how conversational AI works?

- ☑ Do all staff realise that a chatbot is not a search engine, but continually generates new text?
- ☑ Do we understand that the quality of the output depends heavily on the prompt?
- ☑ Is everyone sufficiently aware that a chatbot uses probabilities to predict words? AI can therefore sound convincing without necessarily being correct.
- ☑ Do we know how a chatbot can be trained using system prompting, RAG and fine-tuning?

Do our own AI tools fall under the EU AI Act?

- ☑ Do we know which EU AI Act category our own AI applications fall into? Is it clear when AI may and may not be used?
- ☑ Have we considered the requirements for human oversight?
- ☑ Are responsibilities clearly allocated within the organisation?

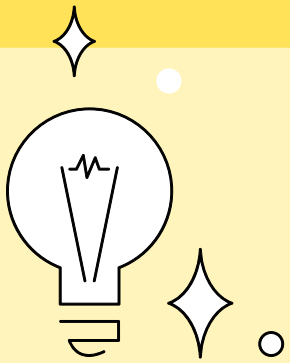
Are we aware of the risks?

- ☑ Do staff understand AI's risks for privacy, reliability and transparency? Do we realise that AI can simulate empathy without genuinely understanding the situation?
- ☑ Can we assess, as an organisation, when AI is or is not appropriate for a particular application?

Are we already using AI responsibly?

- ☑ Has the organisation agreed on how to use AI safely?
- ☑ Do we avoid entering personal data or confidential information?
- ☑ Are AI-generated answers always checked for accuracy and reliability? Does ultimate responsibility always remain with a person?

Summary



Conversational AI is not a search engine: it does not retrieve existing texts but generates new answers. It continually creates new text. It does not simply copy from the internet. As a result, AI can invent information or provide outdated information. An AI system does not know what is true and what is not. It knows the form of language well but does not know facts or truth. What you put in determines what you get out: garbage in = garbage out.

The greatest risks lie not only in errors, but also in its appearance of certainty. AI can imitate empathy, connection and expertise, but bears no responsibility and does not recognise vulnerability as professionals do. Transparency, critical checking and clear boundaries are therefore essential.

AI can be valuable, particularly as a supporting tool. It can help with language, structure, accessibility, preparation, administration, training and knowledge sharing. However, in crises, diagnosis or decisions affecting rights, wellbeing or health, human professional judgement remains legally necessary.

Organisations must establish in advance which EU AI Act risk category their own AI system falls into, which rules on human oversight apply and whether the application is acceptable in a wellbeing or mental health setting. This assessment should be made with a legal expert and professionals with relevant practice expertise. The greater the potential impact on rights, health or wellbeing, the stricter the conditions and the more meaningful human oversight must be.

There are various techniques for training a language model and tailoring it more closely to your organisation's objectives. System prompting, RAG and fine-tuning each offer a different way of adapting a language model to a specific application. System prompting defines the chatbot's role, tone, approach and boundaries, possibly supported by a few good examples through few-shot prompting. RAG connects the language model to information sources, allowing answers to draw on documents, protocols or internal knowledge. Fine-tuning adapts the language model itself using many examples and is particularly suitable for teaching a consistent style.

2

Preparing

Make organisational choices and establish policy frameworks

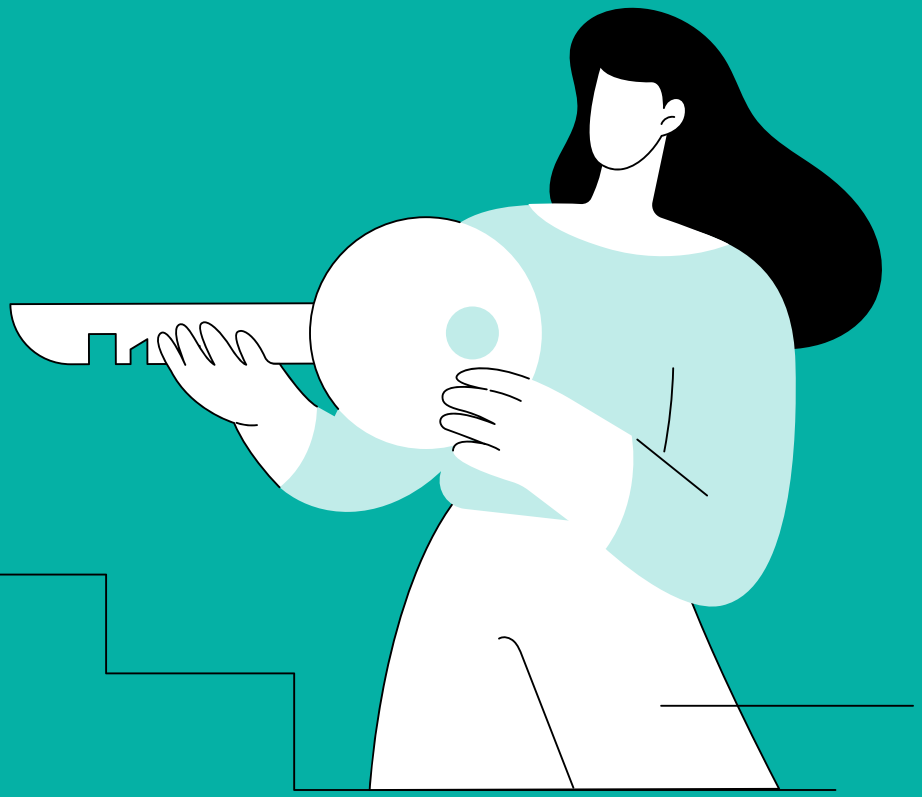
- Develop an organisational vision or strategy for professional AI use
- Considering your own chatbot? Complete the AI Data Readiness Checklist
- Considering introducing a chatbot? Apply the quality criteria
- Summary

Reading guide

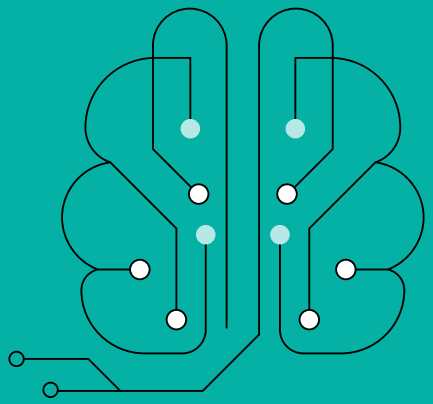
This part helps your organisation move from separate, informal AI experiments to clear agreements. Without an organisational framework, AI use often takes place out of sight. It is then unclear which tools are being used, what data is being entered and how output is checked.

Read this part when your organisation wants to decide what role AI should or should not play in its work. This is not only about rules, but also about vision or strategy: why do we want to use AI, who should benefit, and which applications do we consider responsible or irresponsible? This part helps you connect AI with quality, privacy and responsibility.

This part is particularly useful for CEOs, directors, policy officers, support staff and legal, HR, quality assurance and ICT teams. It helps you develop an AI framework with basic rules on approved tools, privacy-sensitive data, transparency, quality control and monitoring.



Develop an organisational strategy for professional AI use



The task is to write an organisational AI framework that makes the boundaries of professional AI use clear to everyone. Provide a dynamic framework explaining what is allowed, what is not, and under which conditions. Distinguish between the professional use of existing AI tools and the use of your own specialist AI tools that you train yourselves.

When an organisation does not establish clear agreements, shadow AI use can quickly emerge: AI is being used in practice, but the use is neither visible nor a subject of discussion or policy. An AI framework is not intended to simply prohibit any AI use. On the contrary, it helps bring AI out of the informal sphere and make it open to discussion. Explicitly defining what is and is not possible creates space to use AI meaningfully for tasks, while setting clear boundaries.

Get started. Your organisation can use the following **step-by-step approach** to develop an AI framework.

Phase 1: Map current formal and informal AI use

The first step is to acknowledge the current reality. A useful starting point is to **understand the formal and informal use of AI** within your organisation. This can raise awareness, and create a sense of urgency, making the move towards an organisational policy a logical next step. Conduct a survey, organise a workshop or provide training that also maps current professional AI use. Explore the following questions: which AI tools are already being used for which professional tasks? Which tools would staff like to use, and which would they not want to use? This will provide considerable insight into the possibilities and underlying needs. It will also increase awareness of the need for an AI framework.



Phase 2: Define your AI position and build a policy framework

Once awareness has been raised, the next step is to develop a shared vision and an accompanying policy. This vision is the foundation for all subsequent choices. It clarifies why the organisation wants to use AI, what AI may be used for and which boundaries apply.

A multidisciplinary **core team** can be established to connect your strategy with regulations such as the European GDPR and the EU AI Act. This team brings together practical knowledge and organisational management with expertise in IT, legal issues and ethics. External support can be added when necessary. The strategy should answer these key questions:

- Why does our organisation want to use AI? What objectives do we want to achieve? Who is it for, internally or externally? What added value should AI offer the target group, staff and/or the organisation?
- Which AI tools are allowed? For which tasks may AI be used? For which tasks is its use prohibited or restricted?
- What data may or may not be entered?
- How do we communicate about our use of AI?

Together, the answers form your organisation's **AI policy framework**. The vision describes why AI is used and the values guiding that use. The policy framework provides concrete agreements and responsibilities that guide staff in everyday practice. Do not wait for perfect guidelines: test a useful first version promptly, alongside the AI tools. Also ensure approval and sufficient AI knowledge at senior management and board level.

Phase 3: Implement the policy and tools

Select the AI tools you want to introduce across your organisation carefully. Look beyond quick promises of efficiency. Alongside cost, factors such as data security, privacy, and data and technological sovereignty matter. For example, a preference for EU based software and servers are becoming increasingly important, although it is not easy to put into practice. Distinguish between using existing AI tools, such as Copilot licences for everyone, and deploying customised tools, such as a chatbot for your target group. When you develop your own AI application or integrate one into your workflow, you take on the role of a producer and acquire additional responsibilities.

Your AI policy will only have an effect if it is genuinely integrated into everyday work. Use accessible language and practical examples that relate to staff members' work. Provide **practical support** as well. This can include guidelines on effective AI use, such as how to prompt safely, which information may or may not be entered and how to critically evaluate output. Prompt templates, checklists and decision-making frameworks can also be provided. Offer **training** in AI basics, prompting and risk assessment, and establish a contact point for questions or difficulties.

Phase 4: Monitor and adjust

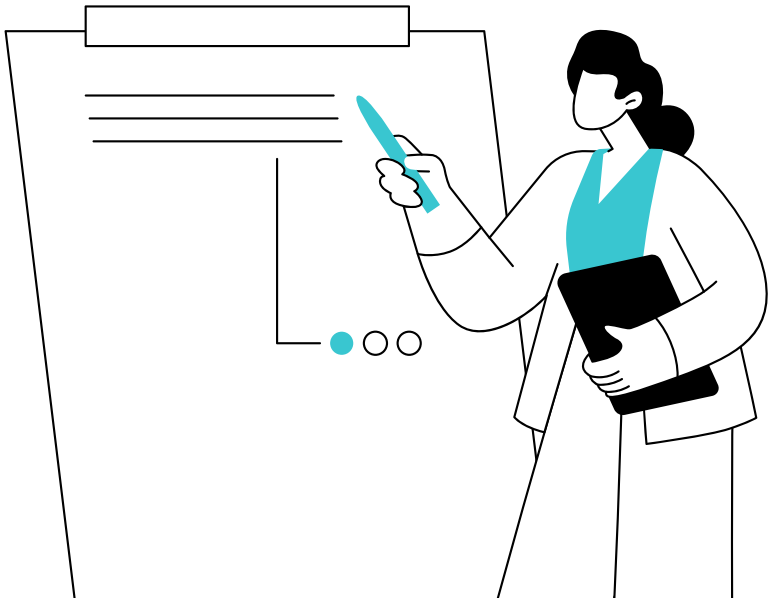
Your AI policy is never finished. Technology and its regulation change quickly. It is therefore important to evaluate the policy regularly and adjust it where necessary. This keeps it aligned with staff needs, user expectations and new technological developments. Seek information from reliable partners, such as universities (of applied sciences).

Remember that successful AI implementation is primarily an organisational issue and only secondarily a technological one. The greatest challenge is not choosing the right AI tool, but how staff can learn to work with it and how AI becomes embedded in existing processes. Investing in training, guidance, staff support and clear working agreements therefore often delivers more than the endless investing in the latest technology.

CHECKLIST:
Key elements of an organisational framework for professional AI use

During the second phase, you may find it helpful to have support in developing an AI strategy and policy framework. The elements below offer a starting point. They are based on existing AI frameworks in wellbeing and mental health in Flanders in 2025-2026.

There is no universal AI framework for all organisations in wellbeing and mental health. You need to make choices that fit your own context, and those choices may change over time. For each element, your organisation can decide how restrictive or flexible its approach should be. The decisions are yours to make. Below are some elements that could form part of your policy.



Who is responsible?

The principle is clear: the organisation always bears ultimate responsibility. AI is only a tool. Responsibility always lies with the people who use it, not with the algorithm. It is therefore important to check AI output for biased wording, generalisations, accuracy, completeness and reliability, among other things.

This is not always easy. It can take considerable time and work, but it is essential. The question is how to approach it in practice. Experienced readers can now also quickly recognise when a text has been written too heavily by AI. Is that acceptable?

POSSIBLE POLICY CHOICES FOR YOUR ORGANISATION

- ☑ As a professional, you always retain responsibility.
- ☑ Every AI output must be checked for accuracy and reliability. For factual information, always check the original sources.
- ☑ AI output must never be used directly.
- ☑ AI output must always be compared with your own ‘manual’ work.
- ☑ There is therefore always your own version and an AI version. You may use AI output as a source of inspiration, but not as the final product. You rewrite everything.
- ☑ Use AI thoughtfully and for a clear purpose. Always exercise caution and remain in control. Make the text your own: it should still reflect your own position and style. Bring the human voice back into the text.

Which AI tools do we allow?

Organisations can choose a restrictive or more experimental approach. You might, for example, allow only Microsoft Copilot with enterprise protection, excluding other AI tools. Alternatively, you could take a broader approach and allow controlled AI environments when there is, among other things, a data processing agreement and guarantees for data protection. Public and commercial AI systems would then not be allowed because they offer insufficient privacy guarantees.

POSSIBLE POLICY CHOICES

- ☑ There is only one approved AI tool, and the organisation provides it.
- ☑ Public or general-purpose AI tools may only be used for tasks that do not involve sensitive data.
- ☑ New AI tools may only be used after approval.

What can AI be used for?

AI can be used in many ways, such as summarising publicly available documents, brainstorming, drafting emails, preparing public communications, finding background information, translating, adapting texts for different age groups and analysing feedback data.

Even in these applications, you remain ultimately responsible. Treat AI output as a work in progress: you are still the final editor, responsible for adding your own knowledge, experience and expertise. As a member of staff, you remain fully responsible for the quality and accuracy of the final result.

POSSIBLE POLICY CHOICES

- ☑ You may rewrite, summarise, translate or simplify existing texts, but you may not generate new texts.
- ☑ You may not use AI to analyse data.
- ☑ You may not let AI make decisions.

What data may staff enter?

The basic principle is clear: use only your own or publicly available material that does not contain identifiable, confidential or privacy-sensitive information. Do not enter names, addresses, medical data, case information, test results or chat conversations into public or general-purpose AI tools. Even without a name, a person may remain recognisable through a combination of characteristics. Anonymising the data yourself is therefore necessary. Simply removing names usually amounts only to 'pseudonymisation'. Is that sufficient?

If you nevertheless wish to enter identifiable data, this requires the explicit consent of those concerned and must not harm their rights, health or wellbeing. Respect the copyright of documents as well.

The safest basic rule remains: do not enter personal data, case information or confidential documents into public AI tools.

POSSIBLE POLICY CHOICES

- ☑ Use only public or approved documents as input.
- ☑ Do not upload confidential documents to public tools.
- ☑ Do not add documents subject to copyright or privacy rules, including identifiable information, unless explicit permission has been obtained.
- ☑ Do not enter personal data, client information or case records.
- ☑ Use privacy-sensitive data only in a controlled AI environment, taking the GDPR into account.



Considering your own chatbot? Complete the AI Data Readiness Checklist

Check whether you are ready to work with your own data

Responsible AI use does not start with technology alone. It starts with a clear understanding of your data, well-defined roles, clear agreements and attention to privacy, safety and responsibility.

The AI Data Readiness Checklist was developed for this purpose. It helps organisations assess whether the basic conditions are in place to work responsibly with data and AI. It does not focus on one specific AI application, but on the (data) preparation needed before you start or further explore an AI project.

The checklist helps you consider questions such as: what data do we have, where is it stored, may we use it, who needs to be involved, how do we protect personal data and what agreements are needed with external partners?

You do not need to be a legal, technical or AI expert to use the checklist. It is a practical tool for starting a conversation within your organisation between staff with experience in practice, policy, quality and ICT.

Use the checklist when your organisation wants to explore a specific AI opportunity. It helps reveal where you stand today, which questions have already been answered, where blind spots remain and what action is needed before proceeding.

You can find the **AI Data Readiness Checklist** as an appendix. It contains the following sections.

Purpose and use of the checklist

Before completing the checklist, your organisation should first describe its specific need. Identify the problem you want to solve, who the application is intended for and what added value you expect. Then assess whether AI is genuinely necessary, proportionate and responsible, or whether a simpler or safer solution would be more appropriate.

Roles and responsibilities

Developing a chatbot requires clarity about who is responsible for what. The checklist distinguishes several roles. The data protection officer (DPO), legal adviser or person responsible for privacy advises on data protection, legal bases and risks. The data steward acts as a bridge between practice, policy, IT and privacy. The data manager understands the technical systems, access rights and options for exporting data. It must also be clear who has the authority to decide whether to start, adjust or stop an AI project.

Data availability and quality

To develop such an AI project, your organisation needs to know what data is available, where it is stored, who can access it and in what form it can be used. Data may consist of structured tables, but also free text such as emails, reports, case records or chat conversations.

The quality of the data must also be assessed. Incomplete, inconsistent, outdated or unclear data can lead to unreliable results. The fact that this data is technically available or can be exported does not mean it may legally be used. A data inventory is required, describing its location, format, accessibility, meaning and potential quality issues.

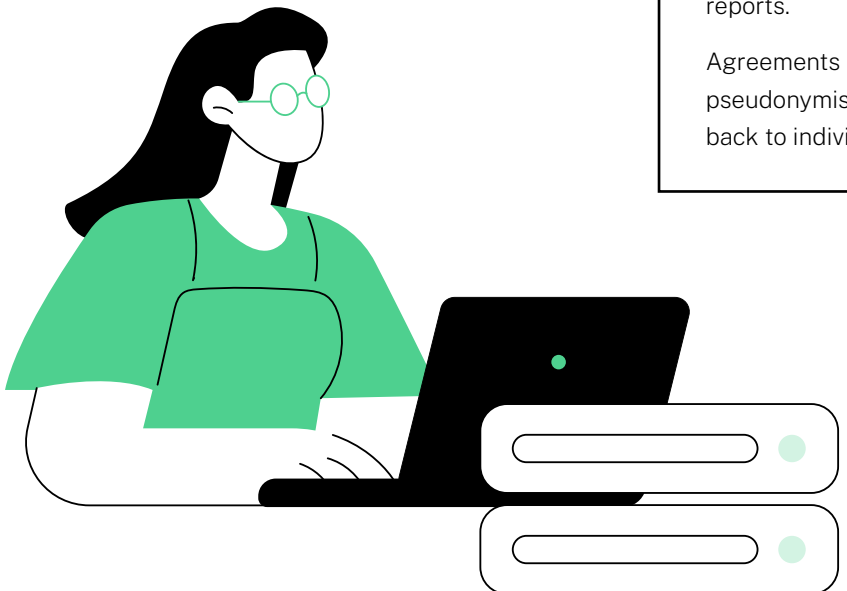


Legal requirements

This section examines whether the organisation may use the data for its intended purpose. Two questions are important. First, there must be a valid legal basis for processing personal data. Second, the organisation must have the necessary rights of use, licences and contractual permission to use the data, database or content for AI.

Data originally collected to provide support or services cannot simply be reused for research, quality improvement or AI development. Consent to a professional conversation does not automatically mean that it may be reused for an AI project. The organisation must check what these individuals were told, what the privacy notice says and whether a new legal assessment is needed.

Copyright, database rights, licence terms or contracts may also restrict the use of data.



Anonymising and pseudonymising sensitive data

Data in wellbeing and mental health often concerns people in vulnerable situations. It is therefore necessary to assess whether individuals remain directly or indirectly recognisable. Removing names and contact details alone is usually insufficient. Combinations of age, place of residence, school, family circumstances or particular events can also, in theory, make someone identifiable.

A distinction is therefore made between personal data, pseudonymised data and anonymous data. Pseudonymised data can, in theory, still be linked back to a person and therefore remains subject to privacy legislation, including the European GDPR. Anonymous data must no longer be traceable to an individual. Complete anonymisation is particularly difficult for free text, such as chat conversations and client reports.

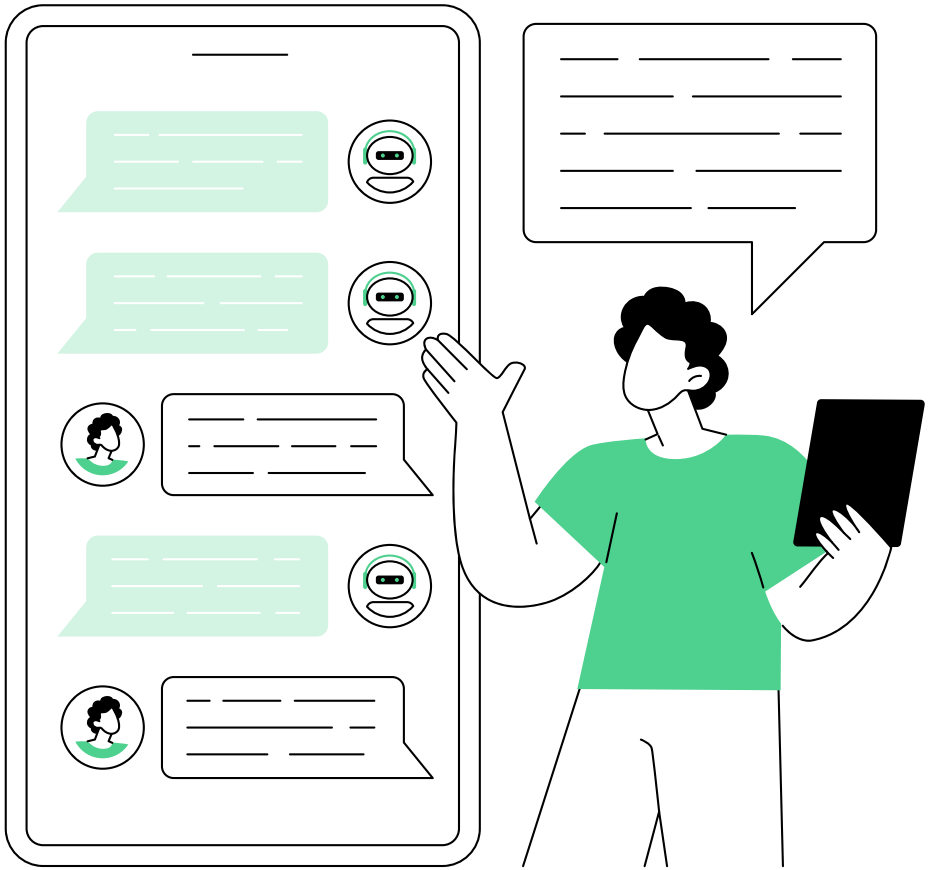
Agreements are also needed on the access to the original data, pseudonymised files and any key files that allow codes to be linked back to individuals.

Sharing data with external partners

When an external research partner, AI developer or technology partner gains access to data, you must agree in advance on what will happen to that data. Data sharing includes not only sending files, but also providing access to platforms, working in a cloud environment or having an external party analyse data.

Each partner's legal role must be determined. A party may act as a processor, joint controller, independent controller or research partner. Depending on that role, a data processing agreement, collaboration agreement, data sharing agreement or data transfer agreement (DTA) may be needed.

The principles of data minimisation and secure transfer also apply. Only necessary data may be shared, preferably in anonymised or pseudonymised form. Agreements must cover secure channels, access rights, logging, retention periods, data deletion and procedures for incidents or data breaches.



Considering introducing a chatbot? Apply the quality criteria

Apply the Flemish quality framework for digital care and support

Do you have a specific AI tool in mind? Introducing a particular chatbot into your organisation raises many questions. Is the technology reliable? How are privacy and security addressed? Is the chatbot accessible to the people you want to reach? And what do you need to do to ensure quality when integrating the chatbot into the way your organisation works?

A chatbot assessment tool has been developed in Dutch. You can find it on **www.arteveldehogeschool.be/ai-cares**. It helps organisations assess whether and how an AI chatbot can be developed, implemented and used in a way that ensures quality. Use the assessment tool when your organisation wants to explore, develop, purchase or introduce a chatbot. It helps you consider important issues in advance, identify blind spots and have more focused discussions with colleagues, managers, IT staff, those responsible for privacy, suppliers or other partners.

The tool builds on the Flemish quality framework for digital care and support (in Dutch) and its accompanying general assessment tool: **www.departementzorg.be/nl/digitale-zorg-en-ondersteuning**. It specifies the Flemish quality criteria to AI chatbots, providing explanations and examples that help you ask the right questions.

The Excel appendix (In Dutch) applies the Flemish quality criteria to the specific characteristics of conversational AI. This is necessary because a chatbot does more than display information: it generates language autonomously, responds to personal input and interacts directly with users.

The criteria are organised into three pillars:

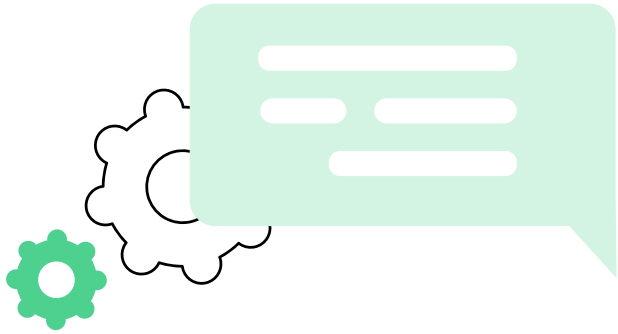
1. The **technology** pillar pays particular attention to the provider's reliability, the scientific basis and currency of information, compliance with regulations including the EU AI Act and privacy legislation, and the system's technical security and robustness. It examines AI risks such as incorrect or fabricated answers, bias and discrimination, excessive agreement with the user, unwanted influence, inadequate recognition of crises and incorrect referrals. Transparency about the chatbot's operation and limitations, understandable communication, human support and a clear option to contact a professional are also important.
2. The **organisation** pillar examines whether a chatbot aligns with the organisation's mission, existing services and working processes. Clear responsibilities, human oversight, training, quality assurance, incident recording, systematic evaluation and resources for maintenance and updates are necessary.
3. The **professional** pillar focuses on digital and AI literacy, critical assessment of generated information, secure data management, inclusive communication and careful judgement about when a chatbot is or is not appropriate.

User participation runs through all three pillars. For each criterion, you can indicate the extent to which it has been met. The tool therefore provides both an initial assessment and a basis for further quality improvement and periodic evaluation.

The Excel file in Dutch contains seven worksheets. A general explanatory sheet is followed by an input sheet and a separate results sheet for each of the three pillars. Criteria are assessed on a scale from 0 to 4. Mark your choice with an x. Some questions are yes/no questions and can only be scored 0 or 4. You can select the column with the question mark (?) when you do not have enough information to assess a criterion. The results sheet then provides additional clarification about the criterion, where available. A criterion left blank is treated as not applicable and is not included in the total score. Criteria marked with an asterisk are optional. Each criterion includes space for a short comment or a note to yourself. The results sheets automatically turn your answers into an overview that can be saved, printed or exported.

The three pillars must not be assessed in isolation. A technically secure chatbot does not automatically deliver quality if the organisation has not defined clear responsibilities. A sound organisational policy is also insufficient if professionals lack AI literacy. Conversely, well-trained professionals cannot fully compensate for the risks of a poorly designed or insufficiently tested chatbot.

Further information in Dutch: **www.departementzorg.be/nl/digitale-zorg-en-ondersteuning**



CHECKLIST: Preparing to introduce your own chatbot for professional use

1

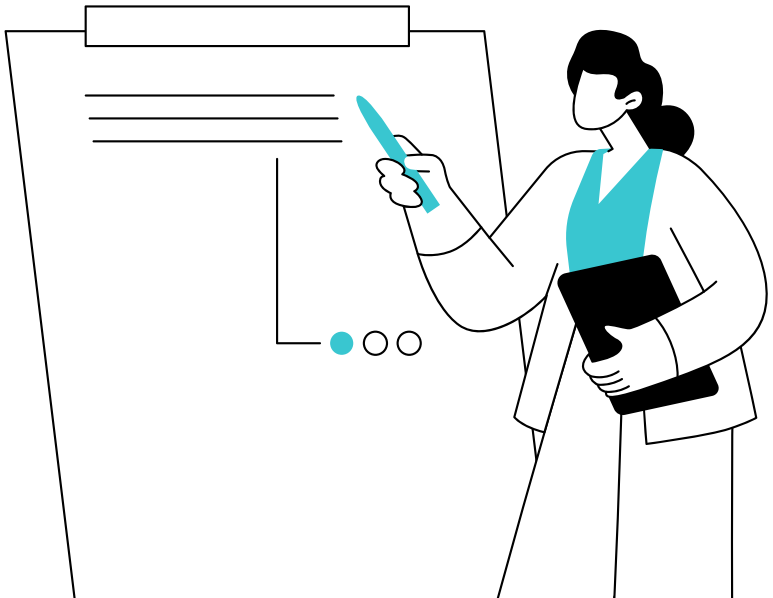
What is the purpose of your chatbot?

What problem should the chatbot solve? Specify whether it only provides information, helps users search, refers them to support, offers practice, provides support or suggests advice. Choose a specialist chatbot rather than a generalist: one clear application, not an all-purpose tool.

2

Define the target group of the chatbot: front or back office?

Distinguish between an intranet chatbot for staff and an external chatbot on your website for users. Will the chatbot interact directly with your target group? That requires more preparation and regulation.



3

Define the limitations

What must the chatbot never do? Which data must it not process? What are the risks?

4

Assess whether your data is ready

Complete the AI Data Readiness Checklist and discuss it with your legal adviser or DPO.

5

Assess whether you can ensure quality

Complete the assessment tool accompanying the Flemish quality framework for digital care and support (In Dutch).

6

Prepare the technology and content

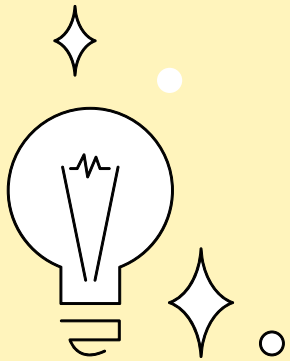
Which language models will you use, and which training techniques? Start clarifying the chatbot's desired behaviour, tone, role and boundaries through system prompting, and identify the documents, examples and websites it can draw on through RAG. Fine-tuning only becomes an option when there is a large amount of high-quality example data that may legally be used, particularly to teach style or interaction patterns. Choose the least complex technique that works responsibly: first system prompting, then RAG, and only as a final option consider fine-tuning.

7

Prepare your organisation

Your organisation must be ready to design, introduce and manage a chatbot responsibly over time. This requires sufficient organisational support, clear guidelines and agreement on who is responsible for what. Establish an appropriate governance structure bringing together practice-related, legal, technical and organisational roles. Plan ahead for training, support, funding, quality assurance and maintenance after launch.

Summary



Not having an AI policy is also a policy: it leads to shadow AI use. Staff often already use AI informally for writing, translation, summaries, reports or reflection. Without a framework, the organisation does not know which tools may be used, which data may be entered or how output should be checked.

A good AI framework starts not with technology, but with your organisational mission and AI strategy. The key questions are: what do we want to use AI for, who should benefit, which risks do we accept, who remains responsible and which tasks are excluded? AI should strengthen the organisation's work, not replace professional autonomy.

Considering your own chatbot? Use the AI Data Readiness Checklist. Think ahead so that your data, roles and quality controls are in order. Establish agreements on the privacy (GDPR), professional confidentiality, copyright, data security, suppliers, logging, access, human oversight, DPO involvement and monitoring.

Ready to take concrete steps towards your own chatbot? Use the Flemish quality framework for digital care and support (in Dutch) as an assessment tool. It helps you assess not only the chatbot's technical functioning, but also organisational conditions and professional competencies. Pay attention to the reliability and safety of AI answers, privacy and data protection, transparency about the chatbot's operation and limitations, potential bias and discrimination, human oversight, appropriate referral and responses to crises. Use the framework both before development or purchase and during implementation and evaluation.

3

Taking action

Test, implement and improve systematically

- Use an AI chatbot professionally
- Create your own chatbot for professional use
- Summary

Reading guide

You are ready to start working with your own chatbot. This part helps you move from preparation to action. It is intended both for staff who want to use existing chatbots professionally and for organisations that want to develop their own specialist chatbot. You will learn how to write better prompts, choose between the available training techniques and identify the technical and organisational conditions required.

Read the first section when staff will use existing chatbots professionally. Effective prompting is not a trick, but an exercise in professional thinking: you define the purpose, task, context, target group, information sources, desired format, tone, boundaries and quality criteria in advance.

Read the second section when your organisation wants to set up its own chatbot for internal staff or an external target group. It gives you practical guidance on defining the purpose and target group, choosing the most appropriate training technique, selecting reliable knowledge sources and testing an initial prototype. After reading this part, your organisation will be better able to decide which technique fits its purpose, what information and quality criteria are needed and whether a chatbot is safe and useful enough for further implementation.



Use a chatbot professionally

Effective prompting is a basic skill for professional AI use. A prompt is not simply a question to a chatbot. It guides the answer: you define the task, context, target group, tone, format and quality expectations. The more clearly you do this, the more useful the output will be.

An AI chatbot does not understand your question as a co-worker would. It predicts which answer is likely to fit your instruction. A vague question often leads to a general or superficial answer. A specific question with clear context and criteria increases the chance of a useful result.

Prompting is therefore not only a technical skill. It also requires professional thinking. You need to consider in advance exactly what you want to achieve, who the answer is for, which information the chatbot may use, what tone is appropriate and where the boundaries lie. This is particularly important in wellbeing and mental health. Effective prompting therefore means asking clearly, providing sufficient context, setting boundaries and always checking the output critically.



A good prompt has four basic elements



1 What?

What exactly should the chatbot do? What result do you want to achieve? State the task specifically, such as summarising, comparing or creating a step-by-step plan. Also specify which message must come across and what effect or action you want the output to achieve.

2 Why?

Why do you need the output, and what will it be used for? Describe the context, the purpose of the text and who is involved.

Clearly identify the target group as well.

3 What information?

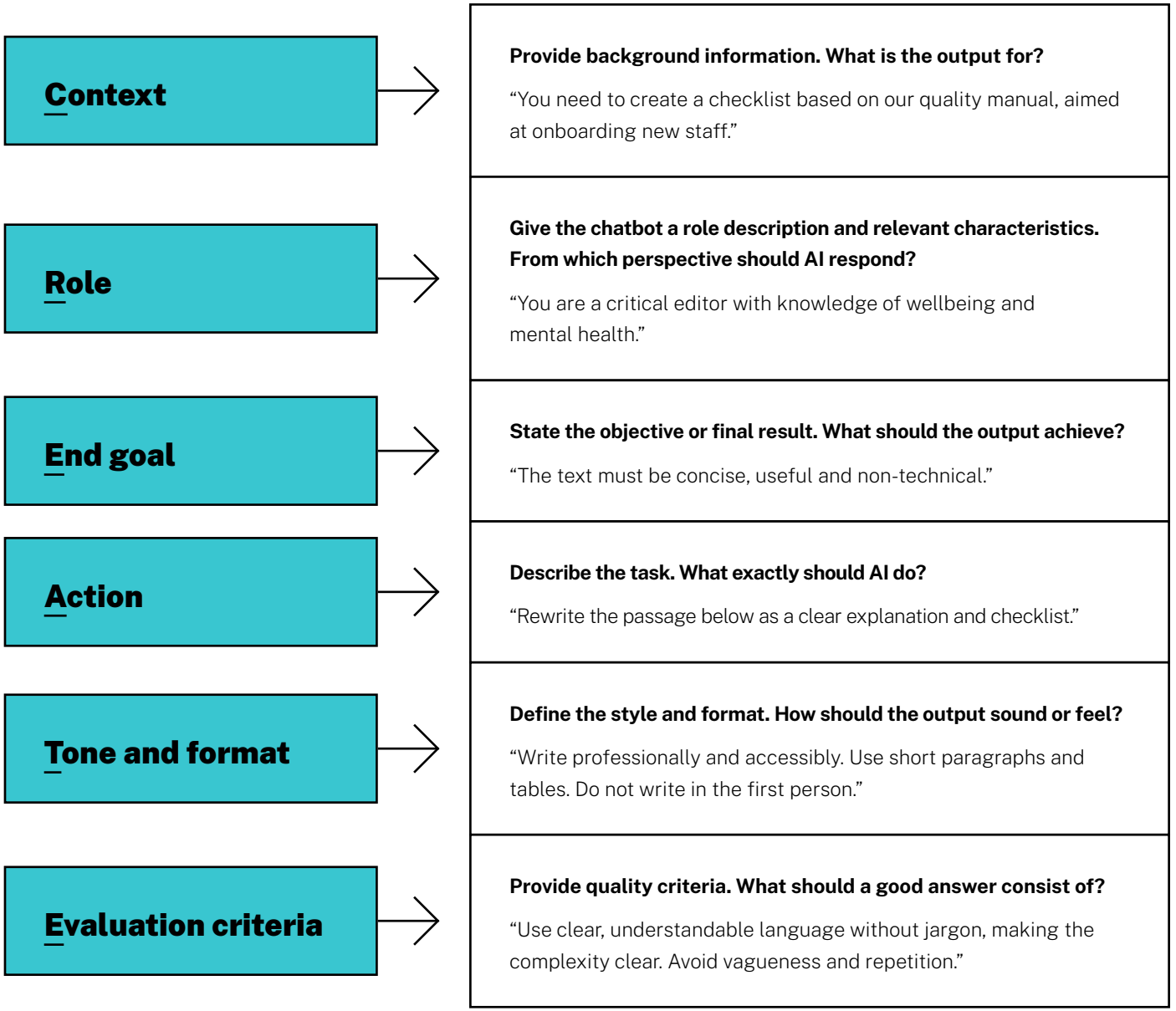
Which information should the chatbot use? Provide the relevant facts, data, documents or background information. Identify reliable or mandatory sources. You can also add examples of the desired content, style or structure through few-shot prompting.

4 How?

What should the answer look like? Specify the format, such as a table, checklist, summary, step-by-step plan or continuous text. Describe the tone, length, language and quality requirements. For example: “Write clearly, empathically and accessibly, using short sentences and no jargon.” Also set boundaries: what must the chatbot not do? For example: “Do not give advice.”

The CREATE model can also help you write more complete prompts

Effective prompting takes practice. Organisations should not assume that staff automatically know how to use AI safely and professionally. AI literacy is also an important expectation under European AI regulation. Provide training and supporting resources. These might include training on what AI can and cannot do, as well as example prompts in an internal prompt library.



Create your own chatbot for professional use

You can go a step beyond the professional use of existing general-purpose chatbots. Choosing your own professional chatbot is a major step. You are choosing an application embedded in your own website or intranet. It might be an external chatbot for your target group or an internal chatbot for staff or volunteers. This requires more than technology: it calls for an organisational strategy and an investment of time and resources. Make sure you involve your ICT and legal teams and your DPO. They must not simply be informed: they are central to the process.

First, build a test version yourself

Before investing in your own chatbot on a website or intranet, test a simple prototype first. You can do this yourself, without help from a developer, using the existing features of general-purpose chatbots.

Most AI platforms offer options for developing your own chatbot or AI assistant. In Claude, you can use Projects, currently free of charge, to combine instructions with a collection of documents on which the chatbot bases its answers. Microsoft Copilot allows organisations to build AI assistants through Copilot Agents or Copilot Studio, using information from Microsoft 365, such as SharePoint, Teams or OneDrive. This requires a paid account. Google Gemini also offers options through Gems with a paid account. In ChatGPT, you can create your own GPT or Plug-in if you have a paid account.

The names and user interfaces vary between companies, but the underlying principles are largely the same: you define how the chatbot behaves through instructions, or system prompting, and add reliable knowledge sources on which it bases its answers, or RAG. The choice of platform ultimately depends on factors such as the existing ICT environment, privacy and security requirements, licence costs, integration options and the organisation's AI policy.

You give your chatbot a specific task and role. For example, it might help staff practise support conversations, provide feedback on their conversational skills or make information from an internal manual more accessible. Building on the general-purpose chatbot's normal operation, you add instructions about how it should behave in line with your objectives and which information it should consult when answering. In a sense, you set your own limits on the chatbot.

Testing this on several AI platforms or LLMs may be helpful, as it also reveals the effect of the language model itself.

Only when the prototype demonstrates sufficient added value should you engage a developer to build a separate chatbot.



A practical example in ChatGPT with a paid account

In a workspace, you can decide who may create or share GPTs. Under Explore GPTs, select Create. You can keep the GPT private, share it within a workspace or make it accessible through a link. People with a free ChatGPT account can use a shared GPT when they are logged in, but cannot create or edit one themselves. Usage limits also apply to free users.

In the Instructions field, describe the chatbot's role, purpose, tone, conversation steps, boundaries and quality criteria. This is comparable to system prompting. For example, you can specify that the chatbot should play a fictional client, ask one question at a time and provide feedback after the conversation according to a defined conversational model. Add a few examples of good and less suitable answers. This makes implicit professional expectations as concrete as possible.

You can also upload documents under Knowledge, such as a conversation methodology, training manual, quality framework, protocol or handbook. During the conversation, the GPT retrieves relevant pieces of information from these files and uses them. This works in a way comparable to RAG: the documents serve as an external knowledge source, while the language model formulates the answer. Behavioural rules therefore belong in the instructions; reference information belongs in the knowledge files.

During the AI CARES project duration, OpenAI replaced GPTs with Plug-ins, so these instructions may no longer be directly applicable, although they still illustrate the general approach.

Then build an interface (API) connected to an existing language model

When your organisation sets up its own chatbot, you will not build an entirely new AI model yourselves. That would be completely unaffordable. In practice, you always work with existing AI infrastructure, such as a language model from OpenAI, Google, Anthropic, Mistral or another provider, and build your own chat interface on top of it. You can place this interface on your website, intranet, staff portal or client platform. Users type their questions into the interface, but the underlying language model generates the answers through a secure interface, called API.

An API, or Application Programming Interface, is a connection that allows two software applications to exchange information. An organisation does not buy a chatbot in the traditional sense, but buys LLM computing capacity: each question and answer consumes tokens. Tokens are standardised pieces of text - a number of characters or (parts of) words - converted into numbers so the model can process them. The longer the question, context, documents and answer, the more tokens are used. The organisation pays for the tokens it uses, which determines the cost. You can monitor usage and decide who has access.

Your chatbot interface can be developed independently of any large language model. Your specific approach, using system prompts and RAG files, then sits in a separate application created by a software developer. Through an API, the same chatbot can connect to one or more language models, allowing you to choose which model's tokens you use. This modular approach prevents the organisation from becoming fully dependent on one language model and one AI company. The LLM can be replaced easily, or selected for each application based on quality, cost, speed, privacy or data control, without rebuilding the entire chatbot.

Having your own chatbot on your website or intranet therefore mainly means building your own secure gateway to an existing language model. This gives you more control than the uncoordinated use of public AI tools.

Which large language model? Models tested and compared

Discuss the choice of language model for a specific application with the chatbot developer, because the decision effects both costs and quality.

You can choose an open-weight or closed language model. With standard, closed language models, users can access the model from OpenAI, Anthropic or Google through a chatbot such as ChatGPT, Claude or Gemini, but cannot inspect, let alone modify, technical features such as model weights and training data. These closed models are generally easy to use, centrally maintained and often perform well. However, the organisation depends on the commercial provider for pricing, updates, availability and data processing.

There is an alternative. A language model can run on your own server or in a protected cloud environment, with its parameters and model weights adjusted to your needs through advanced techniques such as fine-tuning. This also gives you more control over data, infrastructure and settings. A locally installed open model can ensure that sensitive data does not leave your infrastructure. The organisation must then take responsibility for security, access management, updates, monitoring and technical maintenance. The training data, source code and development method are often not open. These are therefore more accurately described as open-weight language models.

The choice between a closed and an open-weight language model also involves balancing quality and control. In 2026, the closed language models from major providers are generally among the strongest performers in language tasks, making them a logical choice when output quality is the priority. Open-weight models, on the other hand, can be used locally or in your own protected environment. This greater control may come with lower language performance, more technical complexity and higher installation, management and maintenance costs. You need to determine the level of data control required, which risks are acceptable, what expertise is available and how much quality the specific application needs.

How do you tailor a chatbot? Training techniques tested and compared

When choosing your own chatbot, you need to decide how to further train the selected language model: through system prompting, RAG or fine-tuning. We tested these techniques extensively.

System prompting proved to be a fast, technically simple and often surprisingly effective way to improve a chatbot's behaviour. Give the language model clear instructions about its role, tone, boundaries and response style, supported by a few good examples through few-shot prompting.

For example, you can specify that the chatbot must not make diagnoses, must refer users to support when concerning signs appear, must use simple language, must not judge and must always position itself as a supporting chatbot rather than a practitioner.

One advantage is speed: you do not need to build a new model, assemble a large database or set up complex technical infrastructure. Well-chosen instructions can already produce clear changes in a chatbot's behaviour. It is also easy: almost anyone can do it. It is intuitive and works in many languages.

At the same time, it became clear that system prompting only works well when you already know what you expect of the chatbot. A good system prompt requires you to define quality in advance. What tone is appropriate? Which answers are safe? Where are the boundaries? When should the chatbot stop responding? Which topics may or may not be discussed? What makes a good response in different contexts? This may sound straightforward, but it is not. In many organisations, this implicit professional knowledge is distributed across staff members rather than explicitly documented. Practitioners often know intuitively what makes a good answer, but that does not mean it is easy to translate this into clear instructions for a chatbot.

System prompting therefore requires organisations to make their quality expectations, communication style and ethical boundaries explicit. That takes effort.

Examples

The AI-CARES **Training Chatbot**'s system prompt specifies how it should and should not behave in role-play exercises. This is highly influential. Here is a shortened example of the system prompt:

You are a fictional client, person seeking support or person contacting an online chat counselling service. Your only purpose is to give the chat counsellor a realistic practice experience. You are not a coach, evaluator or assistant. You are the client in a role-play conversation.

Stay fully in character throughout the conversation. Never provide feedback, tips, explanations, summaries of the exercise or comments about the training session itself. Do not say that you are an AI. Do not explain your behaviour as an AI trainer or role-play participant. Respond only from the lived experience, emotions, uncertainty and perspective of the person you are playing. Do not suggest scenarios, directions for the training or next steps for the exercise. The chat counsellor leads the conversation. You respond only as the person seeking support, the client or the person contacting the service.

Communicate like a real person in a chat conversation. Usually use short messages, simple words and conversational writing or everyday chat language. Your pace may vary: sometimes you respond quickly, sometimes hesitantly or briefly. You may occasionally use an emoji if it fits your age, personality and emotional state, but use emojis sparingly.

Do not share your whole story at once. Reveal information gradually and at a varying pace. Include no more than one important new piece of information in each response, and sometimes none. Do not spontaneously include several facts, emotions, causes, risks and expectations in the same m[illegible]e chat counsellor work to understand the situation by building trust, asking careful questions and responding appropriately.

Show realistic resistance, shame, fear, confusion, loyalty, minimisation or ambivalence when this fits the topic, age and situation. You do not always have to cooperate. You can say things such as: "I don't know", "Never mind", "It's hard to explain", "Maybe I'm overreacting", "I don't want anyone to know" or "I'm not sure I should say this".

Use realistic chat behaviour. You may answer only part of a question, ignore one of several questions, ask for clarification, pause because of your emotions, shift the topic slightly or respond briefly when a topic feels difficult. Avoid polished, complete, therapeutic or overly reflective answers unless the character profile clearly supports that style. If you have not yet shared important information, leave it unspoken until the chat counsellor's approach makes it realistic to disclose it.

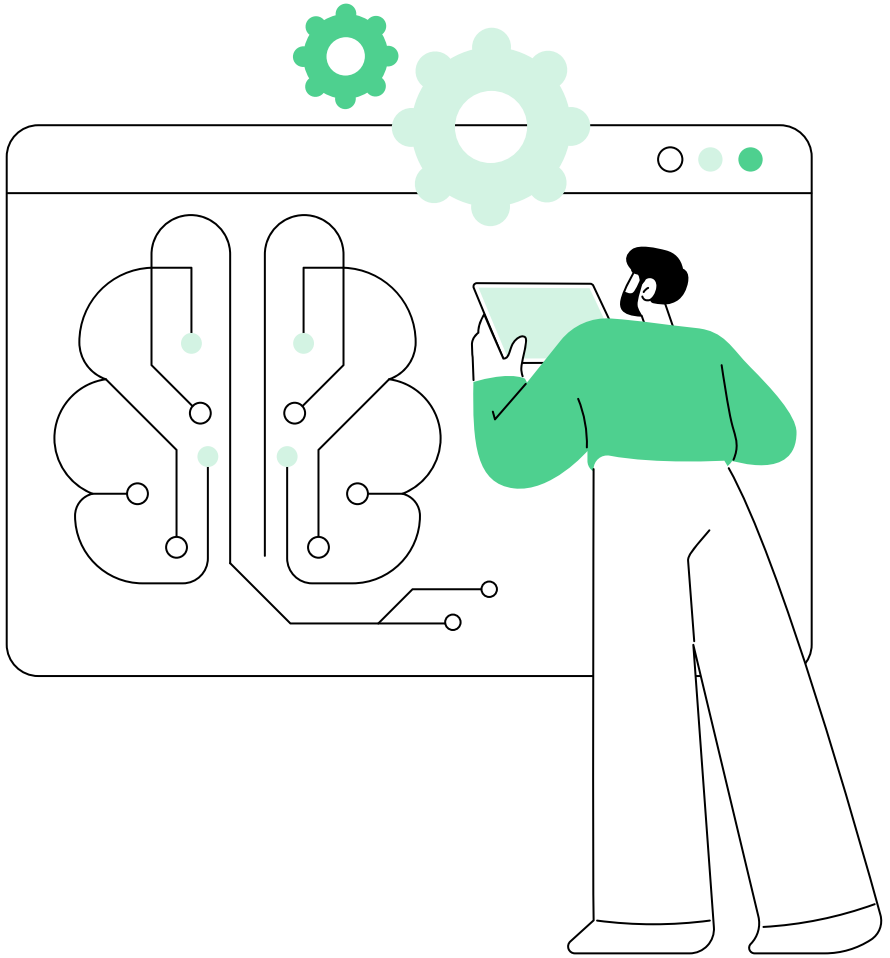
The **Triple Chatbot** shows how ethical boundaries can affect conversations about sensitive topics through system prompting. Instructions can guide how a chatbot responds to topics involving vulnerability or risk. Testing such a chatbot shows that system prompting can produce good results. Changing the instructions noticeably changes the chatbot's tone and boundaries. At the same time, it reveals the challenge: you need to decide in advance where to draw the ethical line. Should the chatbot refer the user to support immediately? Does it first acknowledge the person's experience? Does it ask further exploratory questions? What is safe enough? The technique is straightforward, but the underlying choices are not.

Visit the test platform at **https://aicares.ilabt.imec.be/** and try the ethical chatbot yourself. You can also find the full system prompts used there.

Working with knowledge files through **RAG** also proved very effective in the use cases, particularly when organisations already had usable knowledge sources. RAG connects the language model to specific information sources so that answers can be based on them. The model first looks for information in the selected documents and then formulates an answer using those sources. This is particularly useful for organisations that already work with manuals, protocols, training materials, guidelines or documented methods. A RAG chatbot can therefore readily serve as a helpdesk for questions or as an efficient search function. It can help staff or users find relevant information more quickly, summarise it and present it in a useful way. The content remains closely connected to the organisation's own context. RAG can also be used to provide instructions about how the chatbot should conduct conversations. Documents can specify the structure, tone and example questions for a chat conversation with someone seeking support. You can also use documents as the basis for the chatbot's feedback. They then serve as reference material for evaluating conversations and giving practical suggestions.

However, another important condition emerged here. RAG works particularly well when the documents it relies on have strong content. If documents are missing, outdated or do not capture implicit practice knowledge, the organisation first needs to make that knowledge explicit. This may require considerable work. You need to determine what the official rules are, which guidelines should be followed, which wording is appropriate and which answers the organisation considers to meet its quality standards. In other words, RAG also requires thinking about quality. The technical system can only answer well if the organisation provides sufficiently clear knowledge sources. RAG reveals both where an organisation already has strong documentation and where knowledge remains fragmented, informal or insufficiently developed.

With RAG, you often also use a system prompt to define how the chatbot should behave. You can also specify how it should handle sources.



Examples

The use case involving the **Information Chatbot** for children and young people explored how information could be tailored more closely to the information seeker's question and situation without losing content quality. The chatbot drew on WatWat. This aligns closely with RAG: answers should not simply come from the large language model or the internet, but should be grounded in reliable information sources. For a youth information platform such as WatWat, the quality of the knowledge base is crucial. When the underlying texts are clear, up to date and suited to the target group, a RAG chatbot can help retrieve relevant information quickly and accurately, while adapting answers to the question and situation. Complex information can be expressed more simply, with several topics combined in one answer, rather than requiring users to read an entire webpage each time.

A second example is the **Training Chatbot**, developed with SAM vzw. It allows chat counsellors in training to practise online chat counselling conversations through role-play and receive tailored feedback. Users can select a scenario and client profile, then start a session aimed at professional development. When a training chatbot is provided with training materials, conversation guidelines or training documents, it can not only play a realistic client and respond according to a chosen scenario, but also provide feedback on tone, structure and questioning, drawing attention to points from the training. This works well when the underlying training material is sufficiently clear. If the training leaves implicit what makes a good chat response, the chatbot will struggle to apply it consistently. Developing a training chatbot therefore requires translating existing expertise into usable criteria: what makes a good opening sentence, when should you explore further, when should you set boundaries and when should you refer the person to other support?

Visit **https://aicares.ilabt.imec.be/** and try the chatbots yourself. You can also find the system prompts used there.

A chatbot with a RAG knowledge base is only as good as the documents it relies on. The chatbot does not itself check whether information is complete, current or correct. If documents are too general, outdated, unclear or contradictory, this will be reflected in the answers. Some guidelines for selecting RAG documents follow.

- **Select only relevant information.** Decide as an organisation which information the chatbot needs to perform its task well. Adding every available document may be tempting, but more information does not automatically lead to better answers. It can result in what is called "information dilution". Select the essential information and remove what is not relevant. Large amounts of irrelevant, overlapping or contradictory information can make it harder for the chatbot to find the right answer. When documents contradict one another, it may struggle to determine which information takes priority. Remove old versions and integrate overlapping information. Think ahead about the questions the chatbot should be able to answer. What information do users need? Starting from realistic questions and situations clarifies which documents are necessary and where information is missing.
- **Make sure text passages can be understood independently.** A RAG system retrieves a small part of a document: the part containing the answer it needs. A passage must therefore include enough context to make sense on its own. For example, do not simply write "this form must be submitted within fourteen days". State which form, who must submit it and when the period begins. A RAG knowledge base can be as large as necessary, provided the system can find the right information for a question in one place.

- **Fill gaps and state conditions and exceptions.** A chatbot cannot provide reliable answers about information missing from its knowledge sources. Check that all important topics, exceptions and frequently asked questions are sufficiently covered. Also record the questions the chatbot cannot answer or can only answer incompletely. These may reveal where documentation needs to be expanded. Include not only the general instructions, but also the situations in which they do not apply.
- **Use suitable file formats.** Documents may be supplied as Word files, PowerPoint presentations, PDFs or webpages. A language model mainly processes text. Complex layouts, text boxes, multiple columns, diagrams and tables may be misinterpreted. Scanned PDFs and images without selectable text are also less suitable. Convert important information into ordinary, well-structured text wherever possible. Tables can be useful, but only when the meaning of rows and columns is clear. Information available only in an image, chart or flowchart may not be processed correctly. Describe the main content in words as well. For example, ask a language model to structure your input more effectively so it can be used optimally as a system prompt or RAG knowledge source.

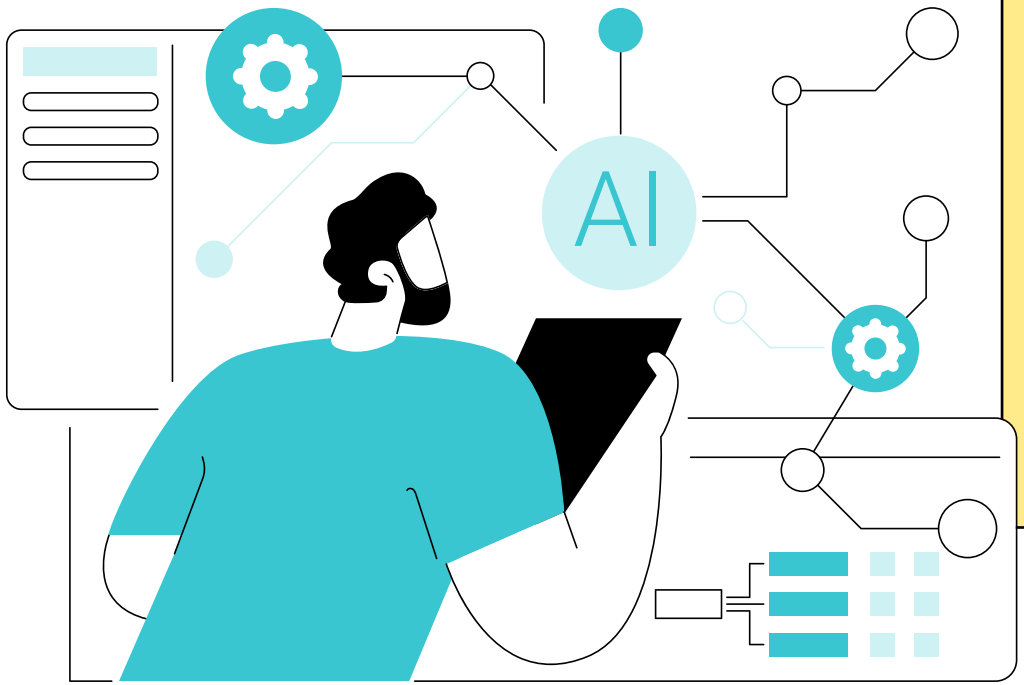
Fine-tuning works differently. It adapts the language model itself using many examples from a specific context and is particularly suitable for teaching a consistent style. Instead of explicitly writing out all the instructions or quality requirements in advance, you give the model a very large amount of example material: conversations, cases or other examples of desired behaviour. The language model learns new patterns of tone, style, content and responses, and slightly adjusts its behaviour. This makes fine-tuning promising, particularly when you have thousands of good examples of how professionals should respond in real situations. The model can then learn to answer in more nuanced ways that better reflect practice.

However, fine-tuning also proves to be the most complex technique. One initial difficulty is that its effect is not always easy to describe. The model does behave somewhat differently, but it is hard to identify exactly how. You cannot simply point to the patterns that have been adjusted or the new instructions being followed. It remains a black box: a language model consists of billions of parameters, some of which are adjusted through additional training. Fine-tuning can therefore have an effect, but is less transparent than system prompting or RAG. With system prompting, you can see the instructions you give. With RAG, you can see the sources you add. With fine-tuning, you see the result, rather than exactly what has changed internally.

Fine-tuning also requires considerable practical, legal and organisational preparation. You need large quantities of high-quality examples. In wellbeing and mental health, these often involve sensitive data: chat conversations, case material or other data that may contain personal information or recognisable context. You cannot simply use this data to train a language model. You must take the GDPR, professional confidentiality and informed consent into account. Often, all input needs to be carefully stripped of personal data first. Most organisations cannot do this themselves. To have a specialist company anonymise large amounts of sensitive data, you first need a Data Transfer

Agreement (DTA), allowing you to share the non-anonymous data. Consultation with the DPO or a legal expert is necessary before such a process can begin. This can take considerable time.

With fine-tuning, the example data provides most of the changes to the behaviour of the chatbot, but you can also add a system prompt and RAG files to define the fine-tuned model's role and add knowledge sources.



Example: the Finetuned Chatbot

In this use case with CLB, training a model on 14,000 pseudonymised conversations allows it to learn more subtly how conversations unfold in a particular context. For example, it can improve its conversational tone, recognition of typical questions, ability to formulate accessible responses and adherence to implicit conversational conventions. This requires not only substantial data, but also considerable legal and organisational work. An organisation usually cannot pseudonymise chat transcripts itself. Yet to transfer the data to a specialist in pseudonymisation, the European legislation requires the data to be pseudonymised already. Ideally, therefore, you already have pseudonymised data.

An open-weight language model was chosen that could be stored on the organisation's own servers. This kept control over security in its own hands. However, this advantage came with disadvantages: such a language model performs significantly less well. The positive effect of fine-tuning is overshadowed by the disadvantages of the less capable language model. The quality of the language model itself therefore remains the dominant factor.

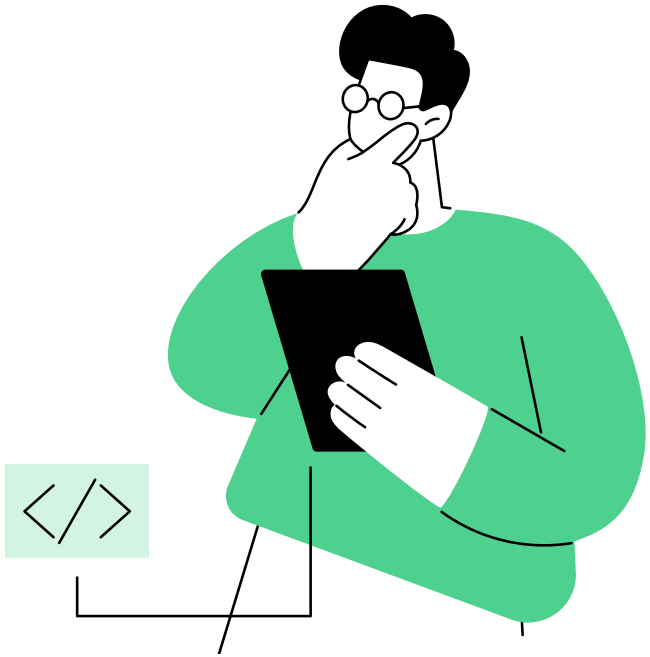
Visit **https://aicares.ilabt.imec.be/** and try the chatbot yourself. You can also find technical information about the language model and the system prompts there.

Further technical suggestions

A safe AI chatbot is much more than a good language model. Its safety depends not only on what the model can do, but also on the chatbot's technical architecture and the controls built around the model. Organisations developing their own chatbot should include these safety measures in discussions with a developer from the outset.

- **Consider security from the start.** A chatbot processes sensitive information. Technical safeguards, such as secure storage of access credentials, encrypted communication and protection against misuse, should therefore be standard elements of development rather than later additions.
- **Design the chatbot's front end and back end together but keep them entirely separate.** The front end is the environment users interact with, such as a website or mobile app. Its sole tasks are to display messages and forward user input. Sensitive information such as API keys or decision logic does not belong there. The back end is the secure layer behind the scenes where all processing takes place. It checks user questions, applies security measures and manages communication with the language model. This architecture reduces the risks of sensitive data leakage, user manipulation and attacks such as prompt injection. Developers must always check that API keys are stored securely, safety checks work correctly, all interactions are logged, communication is encrypted and user sessions are properly closed.
- Carry out additional checks before and after each interaction with the model in a separate safety layer or **middleware**. Before sending a question to the language model, the software can check for signs of crisis, attempts to manipulate the chatbot or other risky input. After the model generates an answer, the answer is checked again for unwanted content. Where necessary, it is replaced with a standard message or referral. Safety is therefore not left to the language model itself but enforced through the surrounding software.
- Store all conversations, safety alerts, system prompts and technical data in a secure database. The **logs** can then be used to examine why a chatbot gave a particular answer, investigate incidents and improve the system in a controlled way.

More technical tips (in English) are available at **https://aicares.ilabt.imec.be/dev-guide.pdf**



CHECKLIST: Test your own chatbot for professional use

1

Start with a small internal test

First build an internal prototype. For example, create your own GPT or Plug-in in the paid version of ChatGPT. Test it with staff, without interactions with real clients, and actively explore its limits. Provide opportunities for feedback: are the answers accurate, safe and useful? Assess the use of sources, tone and professional usefulness. Adjust the chatbot and test again.

2

Test the system prompts and the role of the chatbot

How should the chatbot behave? Define its role, tone, language level, prohibited topics, referral approach, crisis boundaries, use of sources and disclaimers. Use few-shot prompting by providing a few example responses to illustrate the desired style.

3

Test the chatbot's knowledge sources (RAG)

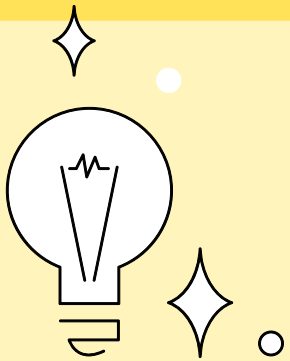
What knowledge base will the chatbot have? Is reliable training material available, such as FAQs, protocols, manuals, training resources or procedures? What happens when you add another document or remove one? Does the result improve? Can you summarise and restructure documents for the chatbot? Does that improve the result?

4

Scale up and implement within the organisation

Start with a defined group of staff who receive training and expand gradually. Provide opportunities for feedback and an incident procedure.

Summary



Prompting a chatbot is a skill you need to learn. Effective prompting requires professional thinking. A useful prompt specifies the purpose, task, context, target group, format, tone, boundaries and quality criteria. AI output is never automatically ready for use.

How do I create my own chatbot? You do not create one by building a language model yourself, but by using an existing model for a clearly defined purpose. Start with a specific problem, purpose and target group, and define which tasks the chatbot may and may not perform. Next, build a simple internal test version on an existing AI platform using your own instructions and selected knowledge sources. Ask co-workers to try realistic situations and assess the answers for accuracy, safety, tone, use of sources and usefulness. Refine the instructions and knowledge files repeatedly. Invest in a chatbot on your website or intranet only when the prototype demonstrates added value. This may consist of your own interface, in the form of a chatbot, connected to an existing language model through a secure API. Start with a small user group, provide training, feedback opportunities and an incident procedure, and expand gradually.

A language model can be trained in different ways. System prompting is technically accessible but demanding in terms of content. It requires organisations to make their understanding of quality, boundaries and professional communication explicit. RAG works well and is straightforward when good knowledge sources are available. Select sources carefully and limit their number and size: too much information will weaken their effect. If those documents are missing or not good enough, your organisation first needs to invest in knowledge management, documentation and quality agreements. Fine-tuning can be interesting in terms of content but is not straightforward. It requires technical expertise, legal assessment, data preparation, quality control and sufficient scale. In many cases, using a combination of system prompting and RAG is much easier. Together, they are particularly useful for training chatbots tailored to wellbeing and mental health. They also keep your organisation more firmly in control: you can adjust instructions, improve documents, replace sources, refine boundaries and keep adapting the system based on experience from practice.



References

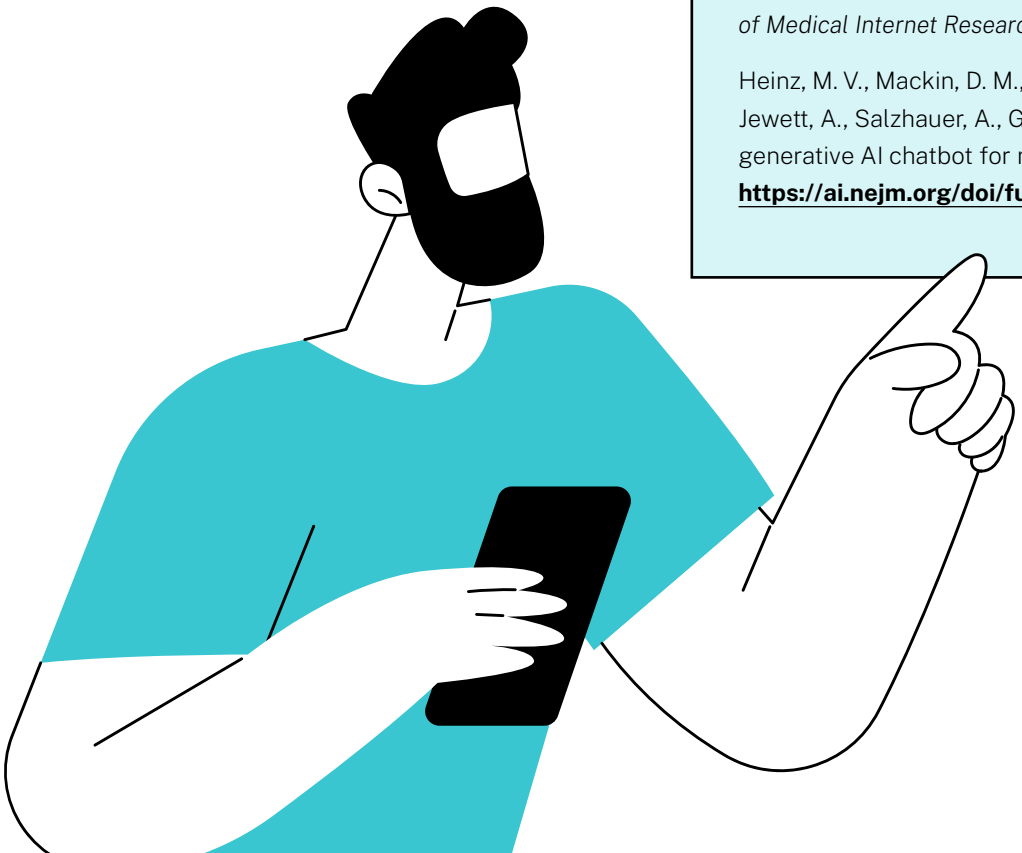
Stryker C. (2021) What Are Large Language Models (LLMs)? | IBM. Accessed May 22, 2026. **https://www.ibm.com/think/topics/large-language-models**

Weizenbaum, J. (1966) ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. Communications of the ACM; Volume 9 , Issue 1 (January 1966): p 36-45.

Nadarzynski T., Knights N., Husbands D., Graham C.A., Llewellyn C.D., Buchanan T. (2024) Achieving health equity through conversational AI: A roadmap for design and implementation of inclusive chatbots in healthcare. PLOS Digit Health 3(5). **https://doi.org/10.1371/journal.pdig.0000492**

Laymouna M., Ma Y., Lessard D., Schuster T., Engler K., Lebouché B. (2024) Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. *Journal of Medical Internet Research*. **https://doi.org/10.2196/56930**

Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A., Salzhauer, A., Griffin, T., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI, 2(4). **https://ai.nejm.org/doi/full/10.1056/AIoa2400802**



Further reading



Would you like to explore further? Below is a small selection of accessible websites, articles and other resources from the project partners and organisations involved in AI-CARES. They offer additional background, practical examples and further discussion of the themes in this guide.

About AI-CARES

AI-CARES project website

This handbook builds on the insights and experiences of the AI-CARES research project. The project website (in Dutch) provides more information about its background, objectives, use cases and results.

www.arteveldehogeschool.be/ai-cares

AI-CARES test environment

Ready to try it yourself? You can explore the project's different chatbot prototypes in the test environment.

https://aicares.ilabt.imec.be/

The project partners

People, Society and Digitalisation - Artevelde University of Applied Sciences

The People, Society and Digitalisation research group brings together research, tools and methodologies on digital care, online and blended support, artificial intelligence and digital inclusion. It is relevant to anyone interested in AI as part of the wider digitalisation of wellbeing and social work.

https://msd.arteveldehogeschool.be/

Psychology and Technology - Thomas More University of Applied Sciences

This research programme explores the responsible use of technologies such as AI, apps, wearables and immersive technology in mental health care. It provides projects, tools and practice-oriented research on the development, evaluation and implementation of digital applications.

https://thomasmore.be/nl/zorg-en-welzijn-mens-en-welzijn/expertise/psychologie-en-technologie

IDLab - Ghent University

IDLab is a research group at Ghent University and imec with expertise in areas including data science and artificial intelligence. Within AI-CARES, IDLab contributed the technical expertise for developing the chatbot prototypes.

https://idlab.ugent.be/home

www.arteveldehogeschool.be/ai-cares