Unit 3 - Limitations and risks of LLMs

Unit 3
Limitations and risks of LLMs

3.1 Unit Introduction

Welcome to the third and final unit of Mastering LLMs!

In the previous units, you have explored LLMs and learned how to interact with them to achieve the desired results.

However, there are times when this isn't possible. In this unit, you will dive into the limitations of LLMs.

You will learn:

- the limitations and risks of LLMs
- outcomes of these limitations

ways to counteract them

Ready to learn more about LLMs? Let's dive into it!

Continue to 3.2: Perfection does not exist

3.2 Perfection does not exist

With AI, it's not always rainbows and butterflies, like the philosopher Adam Levine once said.

While LLMs and GenAI are powerful tools, they have limitations and aren't always flawless.

They depend on the data they've been trained on, don't truly understand things, and can make mistakes, especially in complex or changing circumstances.

Understanding these limitations is important to use LLMs and GenAI effectively, to be able to **set realistic expectations and avoid errors or confusion**.

Think of Harry the Handyman who can fix a leaky tap, but might struggle if you ask him to redesign the electrical system in your house. He can handle most jobs, but when things get complicated, he might not be able to perform the task and might need extra help.

Continue to 3.3: LLM limitations

3.3 LLM limitations

LLMs have limitations because they are algorithms designed to recognize patterns in data, without truly understanding human language or thinking like humans.

There are three main types of limitations.

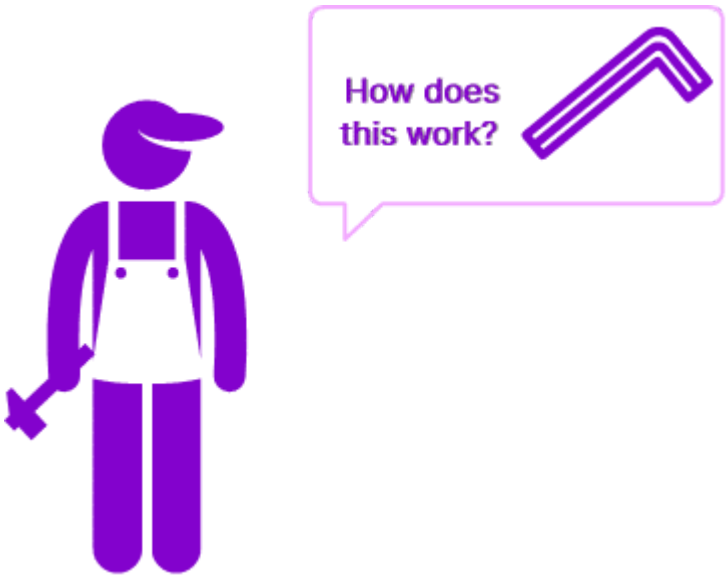
Click each one to learn more.

INTRINSIC MODEL LIMITATIONS | **CREATIVITY AND REASONING** | **DATA PRIVACY AND SECURITY**

LLMs are trained on data and learn how to recognize patterns in it. They **can only work with the information they were trained on** and can't verify facts or adapt to new information that becomes available after the training data was collected.

Moreover, they **lack true understanding** of the text they process. This could lead to outdated or incomplete answers, especially in evolving or complex situations.

It's like Harry who's been trained and is great at using a hammer, but struggles when a job requires an Allen key. He has never seen one before and is trying to loosen the screws by hammering them, not doing a very good job, honestly.



LLMs can simulate creativity by generating responses that appear creative, but they **do not truly create new ideas**. Their outputs are often **recombinations of existing concepts or patterns from the data they've been trained on**.

Harry is amazing at assembling IKEA furniture, but is not able to design a cupboard from scratch.

LLMs often learn from the data users provide in their prompts. If sensitive or personal information is included, the model might unintentionally reuse or expose it later. Without proper safety measures, this can lead to **privacy risks**, like **leaks** or **misuse of confidential details**.

Harry is a bit distracted. While advising a customer on installing a secret safe in their house, Harry gave examples of previous installations but accidentally mentioned a customer's name and the combination to their safe.

Continue to 3.3.1: Intrinsic model limitations

3.3.1 Intrinsic model limitations

You've seen that **intrinsic model limitations** are the **built-in weaknesses of LLMs**.

This includes relying only on their training data, having limited memory, and inheriting biases, which can impact their accuracy and flexibility. Let's have a deeper look at them.

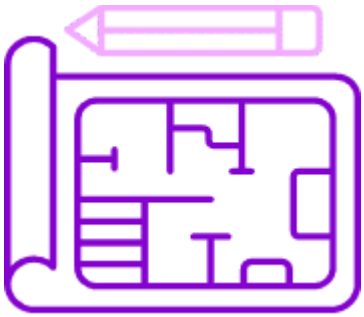
Click each one to learn more.

Knowledge gaps _

LLMs are **trained on a fixed set of data that can become outdated**. Most LLMs cannot access the internet to look for new information or check real-time facts.

LLMs may also **reflect biases present in their training data**, such as social or cultural prejudices, which can impact the objectivity of their responses.

Harry uses old building instructions with traditional materials. He can't check for updates, so he still insulates your building with spray foam, and doesn't know that you can use nanotechnology panels, which are more efficient and sustainable!



Memory constraints

LLMs have a **finite memory** or **context window**, meaning they can only remember and process a limited amount of text at a time. **LLMs forget earlier parts of a conversation** when the limit is exceeded. In long conversations, LLMs may lose track of earlier parts of the discussion, making them less effective at maintaining coherence in lengthy interactions.

Harry can only carry a limited number of tools in his toolbox at once. If the job requires more tools than he can hold, he has to leave some behind. In long projects, he might forget where he put the tools earlier, making him waste precious time and preventing him from completing his job.

Adaptability issues

LLMs **struggle to adapt to new situations and understand abstract concepts**, like justice or freedom. While they can process language and follow patterns, LLMs **don't truly understand the deeper meanings or nuances like humans**. This can lead to issues with abstract concepts or evolving topics, especially when the situation differs from what they were trained on.

Harry can follow a renovation plan step by step but struggles with the finer details, like selecting the right paint color or floor tiles according to his client's taste.

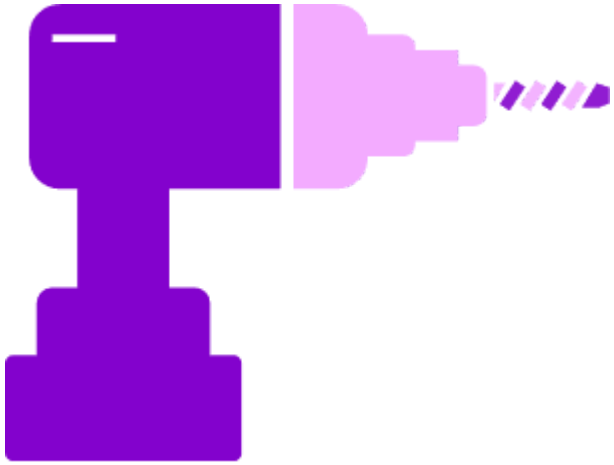
Continue to 3.4: Know the risks

3.4 Know the risks

The LLM limitations you have just seen can lead to incorrect answers, biased results, or security issues if not managed well.

It's important to know these **limitations because they can affect the results you get**. Without this knowledge, you might blindly believe everything LLMs return, which could lead to mistakes or problems. By understanding the limits of LLMs, you can better **evaluate their responses, spot errors, use the technology more effectively and apply strategies to improve their performance**.

In the next sections you will look at some consequences of LLM limitations in more detail.



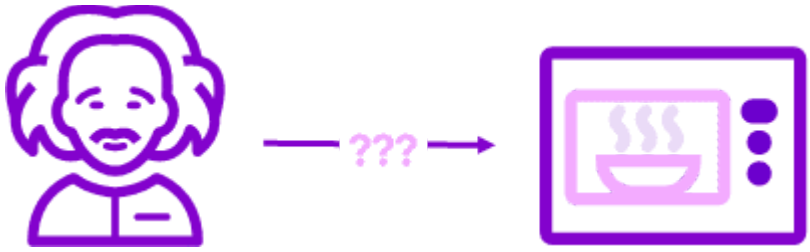
Think of Harry the Handyman, trying to use a drill on a marble tabletop, damaging both the drill and the table. If he understood how the drill works, he could buy the right drill bit, allowing him to finish his job properly and stop causing damage.

Continue to 3.4.1: Hallucinations

3.4.1 Hallucinations

Hallucinations are inaccurate or fabricated outputs.

LLMs can generate outputs that **sound plausible but are completely false or fabricated**. These hallucinations happen because the model doesn't understand the content or the context, it only predicts what might come next based on patterns from the data it has seen.



An LLM might generate a response like for example '***Albert Einstein invented the microwave***', which is clearly false. The model might combine facts it has learned, such as Einstein being a famous scientist and the microwave being a major scientific invention, without realizing the two are unrelated.

Continue to 3.4.2: Lack of robustness

3.4.2 Lack of robustness

LLMs can have trouble staying accurate **when things change quickly** or **when dealing with highly specialized topics** that need deep knowledge or the latest updates.

For example, in fast-changing fields like finance or medicine, new rules, treatments, or technologies happen regularly. If the model's training data is old, it might give outdated or wrong answers. In very specific areas, like advanced physics or complex legal issues, the model might not know enough details and could give overly simple or incorrect responses. This makes LLMs **less reliable in situations where up-to-date or specialized information is really important, especially if there's no human checking the answers**.

Imagine asking an LLM if you should invest in orange juice. If the model uses outdated data, like last season's crop report, it might suggest that investing in orange juice is a good choice. As a result, you might end up buying a lot of orange juice stocks at a time when the value is actually falling.

Continue to 3.4.3: Ethical concerns

3.4.3 Ethical concerns

Click each one to learn more.

Bias

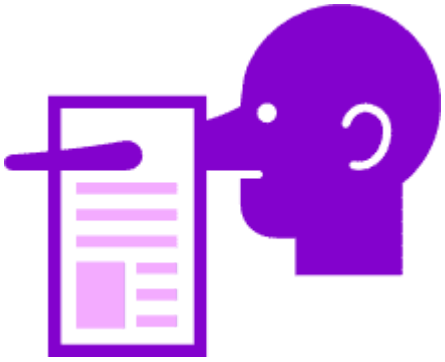
LLMs can unintentionally create harmful outputs because of biases in their training data, leading to **discriminatory language**, **harmful stereotypes**, or **misleading information**. For example, if trained on biased gender, racial, or cultural data, they might unintentionally favor certain groups while excluding others. This could result in biased job descriptions that discourage certain demographics from applying.

Privacy risks

LLMs also present privacy risks. While they don’t store user data, they **might unintentionally reveal sensitive information from their training data**, such as personal details, company secrets, or confidential records **if they haven't been properly anonymized and sensitive information removed**. They could for example reveal parts of a patient’s medical history in a healthcare chatbot.

Misuse

LLMs **can be misused for harmful purposes**, such as **creating fake news**, **phishing emails**, or **manipulative content**. For example, they might be used to impersonate trusted sources, tricking people into making unsafe decisions (like providing bank account credentials) or spreading misleading information.



Continue to 3.5: Counteracting the limitations

3.5 Counteracting the limitations

Not all hope is lost, there is something that you can do!

When you know the limits of LLMs, you can **use strategies to prevent problems and improve the results**.

Imagine Harry is tackling a new project that requires a chain saw he's never used. He doesn't know how to cut through the wood properly. To avoid making mistakes, he reads the manual, watches a tutorial, and asks a more experienced handyman for tips. He's understood his limitations and took action to overcome them. Great job Harry!

Developers use various techniques to improve LLM performance, and **users** (you) can also take steps to address certain limitations.

1: Developers

1

Developers

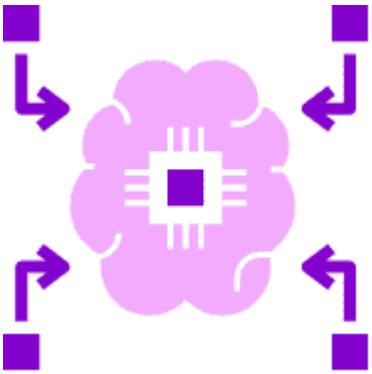
Click each one to learn more.

Retraining

Regularly updating the model with new information to ensure it stays current and relevant.

Memory based extensions _

Enhancing the model with the ability to retain information from previous interactions.



2: Users

2

Users

Click each one to learn more.

Retrieval augmented generation (RAG) _

Integrating external data sources to improve accuracy and provide more relevant answers.

Human-in-the-loop _

Involving human reviews to validate and correct the model's outputs.

Prompt engineering

Crafting inputs in a way that guides the model to generate better, more relevant responses.

Fine tuning

Adjusting the model's parameters using specific data to improve its performance in certain tasks.

Continue to 3.5.1: Retraining

3.5.1 Retraining

Retraining involves regularly updating the AI model with new information to keep its knowledge base fresh and relevant.

This process **repeats the training phase with up-to-date data**, allowing the model to refine its predictions. As a result, the model adjusts the underlying patterns and parameters used for generating responses, making the model more accurate and aligned with current information.

This process helps prevent outdated information and makes it less likely for the model to give irrelevant or incorrect answers based on old data.

Let's learn more about retraining.

Harry is painting a room, but he keeps using the Pantone color of the year from 2021. To avoid being so outdated, he looks up the current Pantone color of the year and starts using it instead.

Continue to 3.5.2: Retrieval augmented generation

3.5.2 Retrieval augmented generation

Retrieval Augmented Generation (RAG) involves the LLM searching external sources, like databases or search engines, for information that's relevant to a user's query.

It **retrieves this data** and then **combines it with the knowledge it already has** to **generate a more accurate, up-to-date response**.

This method allows the model to access current or domain-specific information **without needing to be retrained**, making it more effective for real-time or specialized tasks.

Harry needs to help a customer who got locked out of their house. The customer has a new security lock, so Harry looks up the lock specifications online and opens it easily.

RAG is not automatic in every LLM. The user needs to set it up to get outside information before the model responds.

Continue to 3.5.3: Human-in-the-loop

3.5.3 Human-in-the-loop

Human-in-the-Loop (HITL) means a human reviews and corrects the LLM's output for accuracy.

When the LLM provides an answer, **a person reviews it to make sure it's accurate and makes sense**. If the response is off or incorrect, **the human can correct it**.

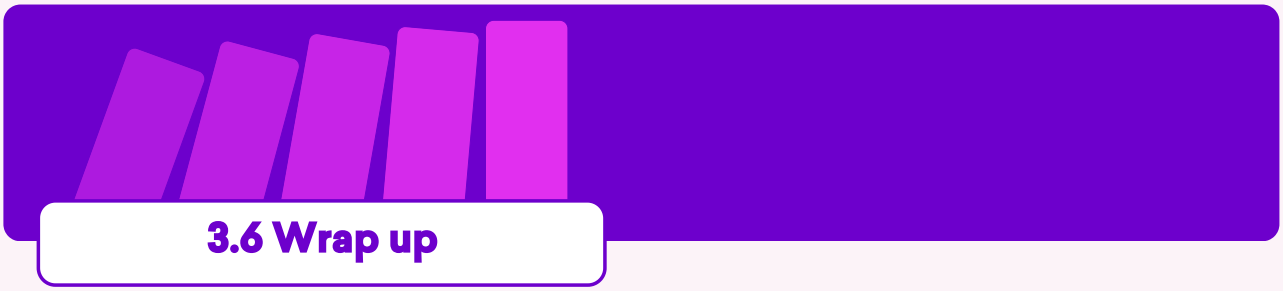
For example, if an LLM suggests a medical diagnosis, a doctor would review the suggestion to ensure it's correct before acting on it.

The **user feedback** also **helps the system get better and avoid mistakes**, especially for tasks that are complex or sensitive.

This process makes the AI more reliable and helps it provide better results.

When tiling the floor, Harry asks for Esther's help. She's a more experienced co-worker and she uses a level to check if Harry's tiling is straight and steps in to correct any mistakes he might have made, leaving the floor as flat as an ironing board.

Continue to the wrap up for this unit



3.6 Wrap up

1

LLMs have some limitations. They **rely on the data they were trained on**, meaning they **can't generate truly original ideas**. They also **struggle with complex reasoning and creativity**. Additionally, LLMs face **issues with data privacy and security**, as they may unintentionally reveal sensitive information.

2

These limits can lead to issues. LLMs might give false or wrong answers, called **hallucinations**. They can also **struggle to adapt to changes and specialized tasks**. Additionally, they can show **biases**, leading to unfair or harmful results.

3

There are ways to address these issues. Developers can **retrain** the model to keep its knowledge up to date. Techniques like **RAG (Retrieval-Augmented Generation)** and **HITL (Human-in-the-Loop)** can **help improve accuracy and reduce errors by using external data or adding human oversight during the process**.

Unit complete!

You've completed the third unit! Now you know that LLMs can make some mistakes and have some risks.

By now you should have an understanding of:

- the limitations of LLMs and the reasons behind them
- the impact that these limitations can have on the generated text
- the techniques you can adopt to mitigate these limitations and risks

Head over to the next course, where you will see some practical GenAI applications and start implementing them in Make.

make | academy

Mark this task complete to continue to the next unit.