

EXTERNAL TOOLS



CONNECT TO OTHER SOFTWARE
SERVICES VIA APIS

MOST USEFUL WHEN

- › LLMs struggle due to their limitations
- › Usecase needs a specialised skillset that LLMs don't possess
- › LLM needs access to internet plugins

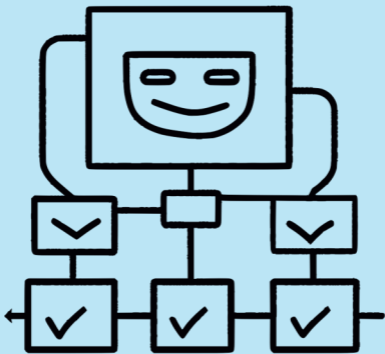
EXAMPLE USECASE

- › Giving OpenAI's GPT4 access to the Wolfram Alpha API makes it way better at performing mathematical calculations.
- › Get access to the current weather in any location using the OpenWeatherMap API.

REQUIREMENTS

- › API access and integration with external or in-house designed tools
- › Instruction tuned LLM
- › Inventory of available external tools

AGENTS



AUTONOMOUS INDEPENDENT AGENTS
THAT EXECUTE CHAIN OF ACTIONS

MOST USEFUL WHEN

- › User goals can be broken down into multiple steps that are consistent repeatable
- › Many of those steps can be textualised

EXAMPLE USECASE

- › Users can tell Google Calendar to schedule appointments with multiple people with different agendas.
- › An LLM then finds out what meetings to schedule, when, with whom and create agendas. And sends that to Calendar.

REQUIREMENTS

- › Integrate LLM with LangChain framework
- › Give access to internet, other applications, if required

BRAINSTORMING



A CREATIVE SPARRING PARTNER TO
GET OUT OF THE CREATIVE BLOCK

MOST USEFUL WHEN

- › Users need creative alternatives and improvements to their ideas
- › Generating a large number of ideas within a short amount of time

EXAMPLE USECASE

- › Explore ideas for social media marketing campaigns.
- › Find synonyms and related phrases to improve writing.
- › Feedback on how some product idea might not work well or could fail.

REQUIREMENTS

- › Good prompt design, tuning tone and attitude to give “constructive criticism”
- › Set a high but reliable temperature for the LLM to create more random output.

SUMMARISING



COMPRESS AND DISTILL LARGE
AMOUNTS OF TEXT

MOST USEFUL WHEN

- › Users need insights from large docs
- › Summaries can be more valuable than the detailed information
- › The devil is not in the details

EXAMPLE USECASE

- › Find key findings from large research publications.
- › Compile simple overview from complicated financial/ legal documents.
- › Find relevant insights from non-fiction books to save time

REQUIREMENTS

- › Access to the data to be summarised
- › Good prompt engineering that conveys intent and context for the summarisation.

EXTERNAL DATA



ACCESSING EXTERNAL DATA SOURCES
TO PROCESS/GENERATE CONTENT

MOST USEFUL WHEN

- › Working with specific domains of data
- › Content needs to be generated based on specific data sources
- › Working in niche/ specialised topics

EXAMPLE USECASE

- › Access different websites/ articles on the internet to find answers, especially about niche topics.
- › Access specific research papers/ reports and answer questions and cite these sources.
- › Refer to a single document for answers.

REQUIREMENTS

- › Internet access for the LLM
- › Prompt design to work on external data
- › External searchable vector database
- › Ability to convert doc to vector database

GENERATE TEXT



COMPOSE TEXT ABOUT AN ENDLESS
VARIETY OF TOPICS & DOMAINS

MOST USEFUL WHEN

- › User activities revolve around writing
- › Writing process needs to become faster or more effective
- › Text generation can be automated

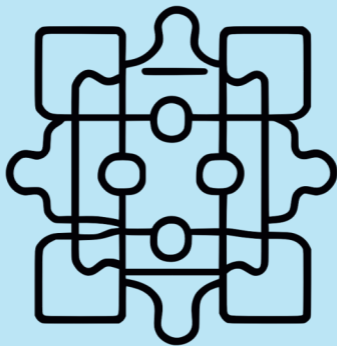
EXAMPLE USECASE

- › Generate emails, text messages tailored to specific people and topics
- › Write articles, blogs, papers using data & arguments about a topic.
- › Write 100 personalised letters instead of 1, addressed to different people.

REQUIREMENTS

- › Access to required specific information
- › Prompt engineering to ensure consistent and repeatable content generation

STRUCTURING DATA



HELP MACHINES UNDERSTAND
HUMANS OR VICE VERSA

MOST USEFUL WHEN

- › When Natural Language input needs to be converted to a standardised format
- › When data or code needs to be re-structured to be easy to understand

EXAMPLE USECASE

- › Users can tell Google Calendar to schedule appointments with multiple people with different agendas. An LLM then figures out what meetings to schedule, when, with whom and create agendas. That can be sent to Calendar.

REQUIREMENTS

- › One-shot or few-shot prompt engineering
- › Fine tune LLM with tabular or structured databases

PERSONAL ASSISTANT



A CHATBOT, BUT TAILORED TO A
SPECIFIC PERSON

MOST USEFUL WHEN

- › Users can benefit from the system remembering information and context about them.
- › Interactions need to be personalised

EXAMPLE USECASE

- › Ironman's J.A.R.V.I.S.
- › Remembers appointments, personal quirks, etc. and takes actions on behalf of Ironman while he saves the world.

REQUIREMENTS

- › LLM needs personal information and contextual understanding
- › Provisions for handling personal information carefully

CHATBOT



MIMIC HUMAN INTERACTIONS AND
COMMUNICATE IN A SIMILAR MANNER

MOST USEFUL WHEN

- › Users need assistance during a process or using a product
- › Products can benefit from guidance or Q&A support

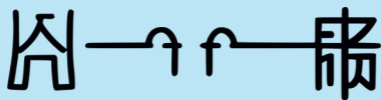
EXAMPLE USECASE

- › Help users with product onboarding and daily use as they figure out different features and tools.
- › Allow users to communicate with their calendar to setup their agenda like Alexa

REQUIREMENTS

- › Instruction tuned LLM
- › Prompt engineering in the backend to give LLM context of the user
- › Testing and tuning to ensure safety

TRANSLATION



因 → ا → f → 膚

TRANSLATE TEXT BETWEEN DIFFERENT
LANGUAGES VIA MULTILINGUALISM

MOST USEFUL WHEN

- › Creating content for multiple demographics
- › Products need to be designed for users working in different languages

EXAMPLE USECASE

- › Hi! Wish you a fun weekend
- › Hoi! Wens je een leuk weekend
- › ¡Hola! Te deseo un fin de semana divertido
- › Salut! Je vous souhaite un agréable week-end

REQUIREMENTS

- › LLM trained on sufficient data from multiple languages
- › Fine tune LLM with text from underrepresented languages

EVALUATING TEXT



CHECKING TEXT AGAINST A VARIETY
OF REQUIREMENTS

MOST USEFUL WHEN

- › Proofreading, spellchecking
- › Tonal analysis
- › Need to write in a specific academic or casual style

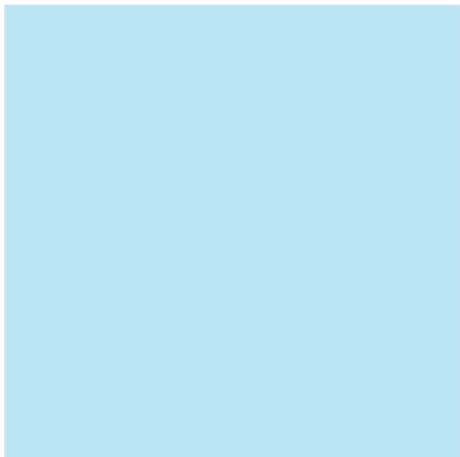
EXAMPLE USECASE

- › Write a draft blog with the relevant info and arguments and ask an LLM to improve the readability, style, grammar etc.
- › Re-write text to make it sound more enthusiastic, supportive, etc.

REQUIREMENTS

- › Well designed prompt from user or backend to ensure consistency
- › Sufficiently capable Instruction tuned LLM

ABILITY NAME



1 LINER EXPLANATION OF THE
ABILITY OR MAYBE 2 LINER

MOST USEFUL WHEN

- › Translating transcripts
- › Referring to specific documents
- › Repeatable multi-step processes

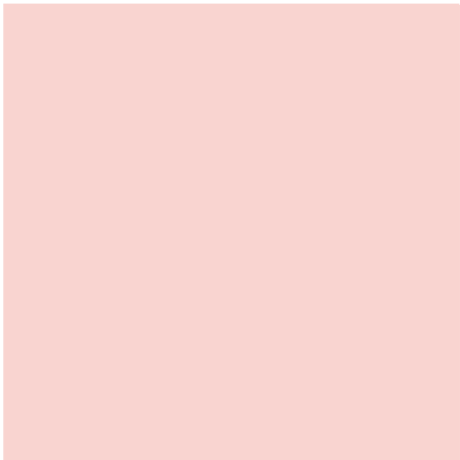
EXAMPLE USECASE

- › Users can tell Google Calendar to schedule appointments with multiple people with different agendas.
- › An LLM then finds out what meetings to schedule, when, with whom and create agendas. And sends that to Calendar.

REQUIREMENTS

- › Hardware, data, code, other efforts
- › Instruction tuned LLM
- › External vector database of text
- › Chain-of-Thought prompt design

RISK NAME



1 LINER EXPLANATION OF THE RISK
OR MAYBE 2 LINER

SAMPLE CONTEXT

- › Sample contexts of when risk might
- › Be likely to manifest
- › Eg. : Biases can emerge when recommending

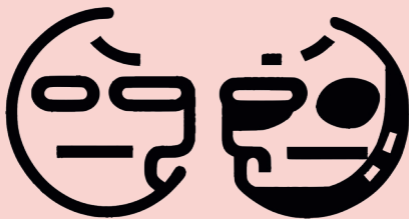
EXAMPLE HARM

- › Detailed example of how a risk might lead to product malfunction or harm to one or more stakeholders.
- › Make example specific and easy to visualise for a broad population.

MITIGATION TIPS

- › Sample of how risk can be addressed via design
- › Or through internal testing
- › Or by monitoring post deployment

BIAS



GENERATED CONTENT REFLECTS
STEREOTYPES AND PREJUDICES

SAMPLE CONTEXT

- › Generating text for sensitive topics
- › Making predictions about people
- › Creating content that might affect public opinion

EXAMPLE HARM

- › LLM application used by a news outlet generates biased and inflammatory headlines that perpetuate stereotypes and incite hatred towards a particular ethnic group.
- › Similar biases also crop up against certain companies, countries & diets.

MITIGATION TIPS

- › Improve & diversify training data
- › Review & test with a diverse group
- › Monitor output for biased content that correlates with specific groups

MISINFORMATION



FALSE AND MISLEADING TEXT,
CAUSING REAL WORLD HARM

SAMPLE CONTEXT

- › Working on news articles, social media posts and scientific publications
- › Situations that need fact-checking and validation to ensure high reliability

EXAMPLE HARM

- › LLM application generates a fake news article claiming a natural disaster has occurred, causing panic and prompting people to evacuate unnecessarily.
- › A scientific article gives incorrect reasoning and causation for findings.

MITIGATION TIPS

- › Use warning label in UI, inform users about limitations of LLMs
- › Monitor output with a robust fact checking process

PRIVACY



REPRODUCE SENSITIVE PERSONAL
DATA FROM TRAINING DATASET

SAMPLE CONTEXT

- › Training or generated data deals with specific individuals
- › Giving personalised recommendations from sensitive data

EXAMPLE HARM

- › LLM application generates responses in a customer support chatbot that contain sensitive financial information about users, exposing it to unauthorized individuals.
- › Similar incidents can happen in the context of healthcare, social services.

MITIGATION TIPS

- › Anonymise & sanitise training data, removing personally identifiable info
- › Restrict access to LLM to authorized personnel only

MALICIOUS TEXT



BAD ACTORS GENERATE HARMFUL,
OFFENSIVE OR SPITEFUL CONTENT

SAMPLE CONTEXT

- › Deploying products on platforms with limited moderation
- › Media generated could be used to deceive or manipulate others

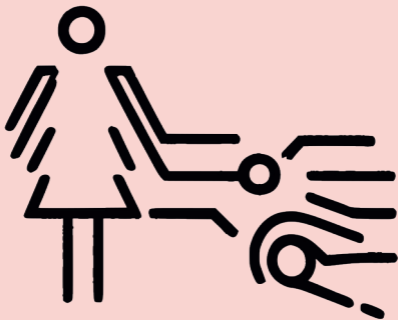
EXAMPLE HARM

- › LLM generates speech for a deepfake video impersonating a political figure, spreading false information and causing public confusion.
- › Similar misuse can include spreading hatred on social media, leading to polarisation.

MITIGATION TIPS

- › Detect & filter malicious content
- › Implement robust content moderation
- › Monitor LLM usage to identify misuse early

OVER RELIANCE



DEPENDENCE CAN REDUCE
JUDGEMENT AND ACCOUNTABILITY

SAMPLE CONTEXT

- › Recommending or automating critical decision making
- › Creating art or public communication without human oversight

EXAMPLE HARM

- › A company heavily relies on a language model to automate customer support responses, leading to numerous customer complaints and dissatisfied clients due to generic and unhelpful replies.

MITIGATION TIPS

- › Avoid anthropomorphisation
- › Design tools that augment not substitute human decision making
- › Involve human review during use

LACK OF CONTEXT



IRRELEVANT OR NONSENSICAL
ANSWERS AND EXPLANATIONS

SAMPLE CONTEXT

- › Long interactive conversations
- › Complex dialogue systems
- › Situations where situational background understanding is critical

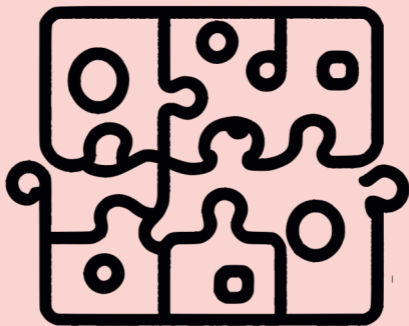
EXAMPLE HARM

- › In a medical chatbot, a language model misinterprets a user's symptoms and provides incorrect medical advice.
- › Sarcasm will be interpreted incorrectly when the LLM does not know the background humour or reality.

MITIGATION TIPS

- › Provide context through prompt engineering or dialogue history
- › Test sufficiently to ensure accuracy
- › Prompt LLM to ask for context

INCOHERENCE



CONTENT IS INTERNALLY
INCONSISTENT & CONTRADICTORY

SAMPLE CONTEXT

- › Creative writing, storytelling
- › Long form content with large context window requirements
- › Arguments need to be built over steps

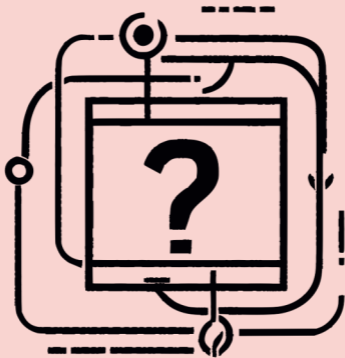
EXAMPLE HARM

- › A language model generates a story with conflicting plot points and confusing character developments, causing harm to the human author's reputation.
- › Similar situation can emerge for scientific or philosophical arguments

MITIGATION TIPS

- › Include training data focussing on long form argumentation and reasoning
- › Test and improve model using RLHF
- › Use prompt engineering to guide LLM

TRANSPARENCY



DIFFICULT TO UNDERSTAND HOW LLMs
MAKE CERTAIN CONCLUSIONS

SAMPLE CONTEXT

- › Products need to be audited for regulatory compliance
- › LLMs are used to support critical decision making processes

EXAMPLE HARM

- › An LLM product denies a loan application without providing a clear explanation, leading to confusion and dissatisfaction for the applicant.
- › Other examples are recommendations for parole, hiring, health insurance, etc.

MITIGATION TIPS

- › Prompt LLM to explain its choices
- › Check training data to understand certain types of wrong responses
- › Explore model agnostic interpretability

JAILBREAKING



DAN MODE, ADVERSARIAL ATTACKS TO
MANIPULATE MODEL BEHAVIOUR

SAMPLE CONTEXT

- › Safety critical applications
- › Online forums, social media platforms
- › Highly valuable/expensive operations

EXAMPLE HARM

- › An adversarial attack causes a language model to produce false medical advice, leading to potential harm to patients following the advice.
- › Microsoft's Tay chatbot on Twitter was conditioned into inappropriate behaviour by other Twitter users

MITIGATION TIPS

- › Integrate robust filtering monitors to detect and prevent adversarial inputs
- › Frequent red teaming to make LLM more resilient to attacks