Document 0 (Source: https://enterthegungeon.fandom.com/wiki/Bullet_Kin)

Bullet Kin

Bullet Kin are one of the most common enemies. They slowly walk towards the player, occasionally firing a single bullet. They can flip tables and use them as cover. They will also deal contact damage if the player touches them.

Occasionally, Bullet Kin will have assault rifles, in which case they will rapidly fire 8 bullets towards the player before reloading. When an assault rifle wielding bullet kin appears, there will often be more in the same room.

On some occasions the player will also encounter incapacitated Bullet Kin lying on the floor. These Bullet Kin are props and disintegrate upon touch. They can be found in mass quantity in Oubliette.

In the Black Powder Mine, they can also ride Minecarts. In fact, if there are any unoccupied Minecarts within the room, they will take priority by walking towards them to ride in.

Trivia

Bullet Kin wield Magnums. Assault-rifle wielding Bullet Kin wield AK-47s.

Incapacitated Bullet Kin can be found in the Oublilette and Cannon's boss room.

In the Oubliette and the boss fight against Agunim, some room props resemble Bullet Kin poking out from inside barrels. This is likely a visual joke on a bullet inside a gun barrel.

In the Portuguese translation of the game, they are known as "Balùnculo", a portmanteau of the words "bala" (bullet) and "homúnculo" (homunculus).

Bullet Kin makes a playable appearance in the platform fighting games Indie Pogo and Indie Game Battle.

Bullet Kin is also a crossover skin in the game Riverbond.

Bullet Kin also has a cameo as lower and upper parts of a skin in the game Fall Guys: Ultimate Knockout.

Veteran Bullet Kin

Veteran Bullet Kin are similar to regular Bullet Kin, but have a higher rate of fire, higher shot speed and attempt to predict the player's movements. They also run faster than normal Bullet Kin, allowing them to catch up with the player quickly if they attempt to take cover.

They fire 4 bullets in a row. If the player moves out of sight from one then the Veteran will pause his attack and then fire the remaining bullets once he has caught up.

Bandana Bullet Kin

Bandana Bullet Kin behave like regular Bullet Kin, but their fire rate is heavily increased. Bandana Bullet Kin also have a higher magazine size than Bullet Kin that wield AK-47s, making them more relentless.

Trivia

Bandana Bullet Kin wield Machine Pistols.

Tanker

Tankers behave like regular Bullet Kin, but have higher health and higher rate of fire. Tankers can be spawned by Treadnaught.

Their rate of fire is slightly lower than that of Bandana Bullet Kin, but they are just as relentless.

Trivia

Tankers wield AK-47s.

The Tanker's expression in his Ammonomicon profile resembles that of the Bullet's avatar when talking to an NPC.

Minelet

Minelets behave like regular Bullet Kin, but will occasionally hide under their hard hat, deflecting incoming projectiles. They will then pop out from underneath their hard hat, releasing a ring of bullets in all directions.

Trivia

Minelets are a possible reference to Mets from the Mega Man series because of their similar behavior. They both hide under their helmets to protect themselves and attack when they emerge.

Cardinal

Cardinals behave like regular Bullet Kin, but have 50% higher health and will occasionally pause to

shoot a group of 5 bullets that will home in on players.

Though a minor effect, these bullets spin around each other as they travel, similar to Apprentice Gunjurers. This occasionally allows them to slip through corners as only some of the bullets will be destroyed.

Trivia

Although normally seen in the Abbey & Hollow, a single cardinal may be seen in the first floor, tending to a small cemetery filled with gravestones. He is the only enemy in that room.

"Of the gun" is a play on the phrase "of the cloth", meaning a member of the clergy.

Shroomer

Shroomers behave like regular Bullet Kin, but have double health and fire two bullets in a V shape. Their bullets can be avoided by standing still, but this can jeopardise dodging the more accurate projectiles of any accompanying enemies. They may also spawn in Gungeon Proper, though rarely.

Trivia

Shroomers will misfire upon spawning, having to stand up after being spawned.

Ashen Bullet Kin

Ashen Bullet Kin have a higher rate of fire and higher shot speed than regular Bullet Kin. They seem to alternate between firing directly at the player and predicting their movements when shooting.

In some rooms of the Forge, Ashen Bullet Kin have the ability to spawn out of ashen statues, which allows them to catch the player off guard.

Trivia

The quote "Cinder Fella" is a clear wordplay between "Cinderella", the famous fairytale, and "Fella" a familiar term for a friend or a person that you consider close.

The French traduction of this quote "Balle au bois dormant" is also a wordplay between the fairytale "La belle au bois dormant" (Sleeping Beauty) and "Balle" (Bullet)

Like its normal counterpart, the Ashen Bullet Kin has a cameo as lower and upper parts of a skin in the game Fall Guys: Ultimate Knockout.

Mutant Bullet Kin

Mutant Bullet Kin behave like regular Bullet Kin, but have higher health and will occasionally stop to release a cone of poison creep. They are immune to Poison effects. The cone of poison can only be released horizontally, so attacking from above or below are the safer options.

Trivia

Its subtitle references Old Faithful, a geyser in Yellowstone National Park.

Fallen Bullet Kin

Fallen Bullet Kin walk towards the player, firing spreads of 3 fire-shaped bullets. They leave behind a small patch of fire upon death. Despite this, they are not immune to fire damage.

Notes

Fallen Bullet Kin will leave their pools of fire in the area where they took the blow that killed them. It will not be spawned where their death animation ends.

Trivia

Fallen Bullet Kin wield Pitchforks.

The sounds that Fallen Bullet Kin make are lower pitched versions of regular Bullet Kin.

These enemies can also be spawned by Lament Configurum.

A portrait of a Fallen Bullet Kin can be seen in the Abbey of the True Gun.

In the Portuguese translation of the game, they are known as "Ex-Balùnculo" (Ex-Bullet Kin), so in that version of the game, it is implied that they are no longer a type of bullet kin, this transformation may have happened through their death, where they were sent to the Sixth Chamber.

Keybullet Kin

Keybullet Kin run away from the player, and drop a key upon death. However, if the player does not manage to kill them in time, they will disappear.

Unlike other Bullet Kin, Keybullet Kin do not deal contact damage if they run into the player.

Jammed Keybullet Kin drop 2 keys instead of 1. These Jammed variations run faster and will take less time to teleport away from the player if they are not destroyed quickly.

If a Keybullet Kin is knocked into a pit, it will not drop a key.

The chances for a specific number of Keybullet Kin to spawn on a floor are:

0 1 2

50% 30% 20%

Trivia

Keybullet Kin may appear in boss arenas during the Boss Rush.

Keybullet Kin have a small chance to appear in elevator rooms at the start of a floor.

Killing 15 Keybullet Kin unlocks the Springheel Boots.

Keybullet Kin and Chance Kin's behavior is modeled after the Crystal Lizards from the Souls series and the Wandering Madness from Bloodborne. Both are harmless "enemies" that quickly run away from the playeroften leading them directly into the path of dangerand despawn after a short time, with the promise of valuable loot if they are killed.

Chance Kin

Chance Kin run away from the player, and drop a random pickup upon death. However, if the player does not manage to kill them in time, they will disappear. Jammed Chance Kins have a chance to drop twice the loot.

The chances for a specific number of Chance Kin to spawn on a floor are:

0 1 2

50% 30% 20%

Trivia

Chance Kin may appear in boss arenas during Boss Rush.

Chance Kin have a small chance to appear in elevator rooms at the start of the floor.

The Chance Kin's subtitle is a reference to the common phrase "No Second Chances."

Chance Kin block player movement during their death animation.

Chance Kin can appear in the same room as a Keybullet Kin.

Keybullet Kin and Chance Kin's behavior is modeled after the Crystal Lizards from the Souls series and the Wandering Madness from Bloodborne. Both are harmless "enemies" that quickly run away

from the playeroften leading them directly into the path of dangerand despawn after a short time, with the promise of valuable loot if they are killed.

Confirmed

Confirmed are mysterious cloaked Bullet Kin. They stroll towards the player, occasionally stopping to fire four slithering lines of bullets at the player from under their hoods.

Confirmed do not appear in specific room layouts. Instead, they have a small chance to replace an enemy in any room. Only one Confirmed can appear on each floor.

Defeating ten Confirmed unlocks the Yellow Chamber.

Trivia

The splash art for Confirmed show them having dozens of red eye-like bullets residing within their cloaks. This bears resemblance to the High Priest's splash art.

The Confirmed are referred to by numerous other names in the game's code, such as 'Kaliber Cultist', and 'Faceless Cultist'.

Red-Caped Bullet Kin

Bullet Kin with red capes will rarely appear in random rooms after at least one Past has been killed. These Bullet Kin do not attack the player, and wander aimlessly. If it is the only enemy remaining in the room and it is left alone for long enough, it will disappear. After this happens 5 times, The Bullet is unlocked, and Red-Caped Bullet Kin stop spawning.

The chances that one will spawn on the six main floors are as follows:

1 2 3 4 5 6

8% 8% 12% 16% 20% 25%

A floor can only contain a maximum of one caped bullet (with one known exception outlined below). There is a 49.95% chance of one or more Red-Caped Bullet Kin appearing in a full run through the Forge, and a 62.46% chance on a run through Bullet Hell.

Trivia

Red-Caped Bullet Kin wield Magnums, but do not fire them or point them at the player.

Red-Caped Bullet Kin do not deal contact damage unless they are jammed.

Red-Caped Bullet Kin's design may be based on The Kid from I Wanna Be The Guy.

Rooms created by the Drill can have a Red-Caped Bullet Kin spawn inside them, even if a Red-Caped Bullet Kin has already appeared on that floor.

It's possible for Red-Caped Bullet Kin to appear in the Aimless Void and Secret Floors such as the Oubliette.

Red-Caped Bullet Kin are not attacked by companions.

Red-Caped Bullet Kin will teleport away if the room contains an enemy that cannot be killed, such as Gunreapers or Dead Blows.

Document 1 (Source: https://www.dropbox.com/scl/fi/ljtdg6eaucrbf1aksw5rm/c2%20-%20session%2050%20-%20underground.docx?rlkey=ioqwgkd14i5xk20i3fp38nzgs&e=1&dl=0)

---The Paths through the Underground/Underdark---(9 days of travel)

Wandering through the dark tunnels, the rushing sounds of the underground river begin to fade as it diverges from the cavern. You walk on for miles, the smell of hard water and wet earth. Natural chambers and cavern passways are chained together by the stretches of burrowed earth left in the wake of this massive worm-like creature. Clusters of crystal and other beautiful minerals occasionally line the walls and ceilings of the chambers, glittering with the little light you have to shove back the darkness.

Day 1 goes without issue... sleep.

Day 2 Ropers

After a few miles of winding tunnel, you emerge in a smaller grotto of stalactites and stalagmites dripping with condensation. Unsure if the same underground river, or another water source, is nearby, you can see quite a bit of ground water does funnel down into this area. Seeking the next burrowed entrance left by the Kryn...

---ENCOUNTER Ropers x 2---

Day 3 goes without issue...sleep.

Day 4 - Kobold Trap

Part way into the journey, the path becomes a protracted tunnel, snaking through the rock for hours without end. Eventually, you begin to notice other smaller tunnels intersecting with the burrowed canal. They appear partially ruined by this fresher tunnel, many of them now filled or partially collapsed.

They are no more than 2-3 feet wide, and numerous (dozens).

In some of the rubble, you can find broken tools... a hammer, some soiled leather, a knife.

The tunnel finally seems to open into a small 15-foot high, 30ft long chamber of dirt and rock, where a rather rancid smell lingers. Glancing within, a handful of the smaller tunnels seem to intersect with it, and whomever enters first (if not Cad), their leg is SNARED by a noose and they must make a Dexterity Saving Throw (DC 15) or be lifted into the air to dangle from a small trap (restrained, DC 16 to escape). The snare also drags a cable tied to numerous pans and metal scraps, making a ruckus!

Chattering and tiny warcrys begin to fill the tunnel from all sides... as dozens of small kobolds rush into the room, and from behind!

-ENCOUNTER: Kobolds x 26, Kobold Inventor x 1-
Loud food! Loud meal!

When seeing the group, they bark and growl. (if noticed, they appear rather fearful)

You! Give us stuffs! Give us foods! Drop things you have, or we stab stab!

If asked about tunnel Big worm eat through! Bring ingoeth! In and out, gone quick, leave mess!

They must parlay with them, avoiding a battle with a significant trade, or intimidation. Otherwise, a fight ensues! Either way, two kobolds are too scared and freeze up. They are brothers Spurt and Bex, scavenger kobolds. They are timid, but know the tunnels well...ish?

https://bytes-and-nibbles.web.app/bytes/stici-note-part-1-planning-and-prototyping)

Semantic and Textual Inference Chatbot Interface (STICI-Note) - Part 1: Planning and Prototyping

The start of my RAG to riches story

STICI-note

Published: Mon, 27 May 2024

Last modified: Tue, 04 Jun 2024

Introduction

In this three-part series, I will be talking you through how I built the Semantic and Textual Inference Chatbot Interface (or STICI-note for short), a locally run RAG chatbot that uses unstructured text documents to enhance its responses. I came up with the name when I was discussing this project with a friend and asked him whether he had any ideas of what to call it. He said, "You should call it sticky because it sounds funny." The name... stuck...

The code for this project is available here.

In this part, I will be planning the project from the tech stack to the techniques I will use, and I will be building a prototype. I will be discussing all of the choices I made and all of the choices I didnt make, so I hope you find this insightful. Without further ado, lets get started.

The Problem

In my spare time, I occasionally play Dungeons and Dragons (DnD), a tabletop roleplaying game, and the stories are often told over several months, so details can be easily lost or forgotten over time. I can write notes on my laptop, but sometimes regular text search does not always provide me with the results I want when trying to search for specific notes. Some common examples include when a keyword is used often (e.g., I might write a lot about the actions of Miles Lawson, but only one segment of text might describe who he is, making searching for information on his character like

finding a needle in a haystack) or when I simply cannot think of the correct keyword to search (e.g., what if I search silver instead of gold?).

One day, I thought to myself that itd be great if I had a tool that I could write my DnD notes into in an unstructured way and retrieve the information at any time with simple questions like Who is Miles Lawson? or How much silver did I pay for a stay in ye olde tavern?. This tool could be extended to be used for querying my notes on many things that are not available online (and therefore not searchable on a search engine), such as documentation on software that I build, notes on things that Im learning about, such as AWS cloud infrastructure, and my diary of my deepest thoughts and feelings (at least I hope this is not available online). And thus, I decided to start working on STICI-note because the tools available online that do this cost money and run on the cloud, and Im a tight-fisted guy whos very sceptical about company promises to not sell your data.

Narrowing Down Features

As with all projects, I began by deciding what features I needed from this tool.

Required features:

Chatbot that you can ask questions and get answers in response (conversational memory is not required).
Information is taken from an unstructured text file.
It must be able to tell me if it doesnt know the answer to my question.
Fast.
Efficient enough to run on my MacBook with other programs without any performance issues.
Locally run for privacy and to ensure it will always be free, runnable, and consistent.
Conversational memory is the memory of previous interactions given to an LLM. I decided not to require it as a feature because I just need the AI to answer my questions about the given text. It might be added as a feature in the future if I feel like I need it, but I do not plan to include it in the initial version of STICI-note.

I knew that limiting it to running on my M1 MacBook with 8 GB of memory would greatly limit the performance of the tool as I would not be able to access truly large language models like GPT-4 and

Claude 3 Opus, but I decided to do it anyway primarily for privacy but also to remove dependencies on external organisations to reduce the maintenance required for the tool in the future.

Planning How to Evaluate and Compare Solutions

If you dont evaluate a solution, how do you know whether its an effective solution? You dont.

I next planned how I would evaluate different variations of the tool. While I do not evaluate anything in this part, I decided to sketch out a rough plan of how I would evaluate different solutions to encourage designing a testable AI in the same way that Test-Driven Development (TDD) encourages you to write testable code.

At first, I considered using Wikipedia pages as the data source and making my own questions about the content of the pages before I realised that this would lead to data leakage as many LLMs are trained on Wikipedia data.

An alternative dataset that I considered using for evaluation is the TREC 2024 RAG Corpus. This is a 28 GB corpus of text documents scraped from the internet. This corpus comes with thousands of queries, and the relevant documents for each query have been tagged as such. This is an amazing corpus for training and evaluating vector DataBase (DBs). Ignoring the fact that its questions do not come with answers, meaning I would have to write my own answers to use the document, there is one glaring flaw that makes it unusable for my use case: the documents are generally relatively short and describe a large variety of things. In my use case, I expect documents to be long and typically written about the same topic. If I were to use the corpus, I would have to stick documents together to present a realistic challenge in the semantic search of the vector DB vector space, but as each document will likely be about very different topics (e.g., one might be about aviation while another might be about economics), context retrieval would be unrealistically easy.

Another alternative evaluation dataset that I considered using was a synthetic dataset. By following a method like this, I can use an LLM to generate synthetic context and questions automatically. I decided not to do this as I was concerned that this would produce bad-quality data with a massive bias towards things an LLM might already know, despite the use case expecting data that the LLM does not already know.

Because the documents in my evaluation corpus need to be thousands of words in length while staying relevant to a topic and they need to include information that will not be in the LLMs training data, I decided that it would be best to manually curate a small dataset to evaluate my models. I plan to create documents from sources on the internet like videogame wikis, peoples blogs, and scientific journals and write my own pairs of questions and answers about them. I will then evaluate the difference between the models answer and my answer using a semantic similarity score.

RAG vs a Really Big Context Window vs Infinite Context

To be able to answer questions about documents, the LLM would need to have access to information from the documents. I thought of three potential solutions for this:

Retrieval Augmented Generation (RAG)
An LLM with a really big context window
An LLM that supports infinite context length
Using an LLM with a really big context window such as Anthropics Claude 3 and Googles Gemini 1.5 would certainly give me the best results as it would allow inference using completely unfiltered and uncompressed context as they can handle inputs of over 700,000 words, but these models are closed-source, and there is absolutely no chance of a model of this size fitting into my tiny M1 MacBook with 8 GB of memory.

By an LLM that supports infinite context length, I mean models like Mamba, Megalodon, and Infini-attention that compress context into a finite space. I decided not to use a model like this for two main reasons. Firstly, I have concerns about the performance. These architectures are in their infancy, and I do not expect them to outperform equivalently-sized traditional transformers. Secondly, as these architectures are very new and experimental, I do not expect much support for them, especially for Apples M-series of chips, which have their own graphics API, metal, that is required for GPU acceleration on my MacBook. These architectures are very interesting, and I would love to try them out, but for this project, I will have to settle for a more tried-and-tested approach.

The more tried and tested approach that I settled with is RAG. It is an incredibly popular technique

for allowing LLMs to make use of information that is too big to fit in their context windows. This technique is known to perform very well, is incredibly well supported by LLM frameworks like LangChain and llamaindex, and works well in resource-constrained environments like on my laptop. Given all this, RAG was an obvious choice.

Optimising Models for Limited Memory Environments

Next, I decided to investigate what kinds of optimisation strategies were available to use to try to fit bigger models into my M1 chip, as bigger LLMs typically perform better (I know, a groundbreaking revelation). To optimise the LLMs that I use, I considered four different techniques:

Quantisation
Model pruning
Model/knowledge distillation
AirLLM

Quantisation is the most common method for making ML models smaller (and therefore faster and more capable of fitting into smaller spaces). Its well known for improving speed and memory usage with little loss in accuracy in return, which makes it very popular for production-level AI. Quantising a model would require being able to fit it into your GPU, but Im trying to quantise a model so that it can fit into my GPU, so without additional computing power, its a bit of a chicken and egg problem. Luckily, because this is such a popular technique, there are many quantised versions of large, high-performance LLMs available on HuggingFace that I can use, so there is no need to do this myself.

Model pruning is a less common method for reducing model sizes, but it is not a technique that one should overlook. This is a technique that can be combined with quantisation (or used on its own) to further reduce models at the expense of accuracy, but I do not plan to apply it myself due to its complexity and the fact that quantisation has the same effect. There are pruned models available on HuggingFace, but they dont typically perform as well as equivalently sized quantised models, so I do not plan to use any unless they have particularly good evaluation results on a common LLM benchmark.

Model/knowledge distillation is another size reduction technique that I considered. Unlike the

previous techniques, model distillation can actually improve accuracy in domain-specific tasks while making a smaller model. As with quantisation and model pruning, I will use pre-distiled models, but I will not distil any models myself due to the computing power it requires (which admittedly is far less than training a model from scratch) and the complexity it would add to the project.

The final optimisation technique that I considered, AirLLM, is quite different from the others in that instead of optimising the model weights, it optimises the model inference. Typically, LLMs are loaded onto the GPU in their entirety, requiring a lot of VRAM to run the larger, better-performing models. AirLLM is an open-source library that tackles this problem by using layered inference, an inference technique that involves loading layers individually when they are needed instead of all at once. This allows larger models to fit into smaller memory spaces without degrading performance. This method definitely has a high potential for accuracy, but I decided not to use it as I am concerned about compatibility and reliability issues as it is a new tool and the GitHub repo has been developed by a single person, so support for it is likely to be limited. Additionally, my M1 chip only has 8 GB of memory shared between the CPU and GPU, which is excellent for reducing data loading overhead costs, but it means that larger models that require AirLLM will be loaded directly from the SSD, so I am concerned that the model layer loading and unloading will become a massive bottleneck when doing inference on larger models. I will reconsider this option if I find that the models that can run on my MacBook do not have satisfactory accuracy.

What Models Even Run on My MacBook?

After getting an idea of what kinds of optimisation techniques were available, I decided to conduct some tests to find out what LLMs would actually run on my MacBook. You could argue that since I am only building a prototype right now, I only need to find one LLM that performs well on my MacBook, but I decided to find five models instead to give me an idea of what kinds of models I will be able to use. In particular, I wanted to know how big the models I could run were and what precision the weights would likely be.

I tested models that I had heard were good or showed decent results on the Hugging Face H4 Open LLM Leaderboard. I found LM Studio incredibly useful for testing out LLMs without having to write any code, which saved me a lot of time. Below are the five suitable models that I found that could run on my MacBook and were fast enough to satisfy me:

tinyllama-1.1b-chat-v1.0 Q6_K

Phi 3 Q4_K_M

bartowski/dolphin-2.8-experiment26-7b-GGUF Q3_K_L

mgonzs13/Mistroll-7B-v2.2-GGU

QuantFactory/Meta-Llama-3-8B-Instruct Q3_K_M

These models range from 1.17 GB up to 4.02 GB in size. I chose not to use any models that were any larger than 4 GB, as with only 8 GB of memory available, I expect that models that are any bigger would seriously impact the other applications that the user (i.e., me) is running on their device.

I will likely test out more models than this while testing out different configurations for the tool, but for the prototype, this is enough.

A Model Without a Framework is Like a Car Engine Without a Chassis

To run my models, I could have written a framework for loading, unloading, and executing the models, passing context and queries to the models, and integrating the vector DB (more on that later) with the inference model from scratch, but I didnt because Im not insane and I am not trying to learn how to make ML frameworks. A lot of university students (myself included) are conditioned to try to build things from scratch for fear of plagiarism and because they are used to building things from scratch as a learning exercise (a very effective one in my opinion), so its difficult to unlearn the DIY mindset, but its simply a lot quicker and a lot more reliable to use libraries than to reinvent the wheel. Saying that, I decided to use a relatively simple tech stack.

Python was an obvious choice for me, given that I have a lot of experience with it and that it has an abundance of support for machine learning applications. I decided to use LangChain to orchestrate my RAG process from the vector DB to the inference, as it is a flexible tool for composing NLP pipelines. It is very popular and reliable, and it includes a lot of tools that make developing NLP applications easier. I considered using LlamaIndex as it is built more specifically for RAG applications, but LangChain is more general-purpose, which I expect will make it more extensible for times when I might want to add more features in the future. Additionally, I am more likely to use LangChain again for other applications in the future, so the experience will be more useful. I also

considered using LitGPT, but I had some issues getting it to work with the M1 chips Metal Performance Shaders (MPS), so I decided not to use it for fear of incompatibility. LitGPT is also intended more for training and fine-tuning LLMs, so it is likely not the best tool for simply deploying them in an application.

To run inference on my models, I will need another library to actually execute the model. As I am using a range of pre-trained models, I will mainly use HuggingFaces transformers library and the Python bindings for llama.cpp library to load and execute models, as these provide simple interfaces for inference, and I dont need the additional control that deep learning frameworks like TensorFlow, PyTorch, and JAX provide as Im using pre-trained models. As I mentioned earlier, AirLLM is still on the table if I need better performance, but I will find out while evaluating models whether this is necessary.

Magnitude + Direction DBs

Since I was using RAG, I needed a vector DB. Deciding which one to use was the final step of the research and also the most difficult one, as vector DBs are the technology that I am least familiar with. For the vector DB, my main requirements were simple: it needed to be lightweight, locally runnable on a laptop, fast, and compatible with my MacBook. Lightweight and locally runnable sound like similar things, but I mean different things by each phrase. The locally runnable one is quite self-explanatory, but by lightweight, I mean quite minimal computation requirements that dont add features like heavy amounts of redundancy and heavy caching, which are useful for large-scale systems, but will simply drain resources in my application that is designed to run alongside an LLM and the users other applications on 8 GB of memory shared between the CPU and GPU.

I considered sixteen different vector DBs, but there were three different solutions that stood out to me for my use case: Chroma, Qdrant, and Vespa. These were all lightweight vector DBs that fit all of my requirements, but I decided to pick Chroma out of the three of them because it has a very batteries included approach, which is very useful for getting an application working quickly.

Below is an overview of my planned tech stack. I did not include any UI, as there are many options for this, and I plan to decide on the UI when building the final product.

The tech stack and data flow that I plan to use for STICI-note.
The tech stack and data flow that I plan to use for STICI-note.

Making the Prototype (A.K.A. the Easy Part)

The final step was to actually build the prototype/proof of concept (PoC) to find out whether this idea can actually be implemented using the tech stack that I have picked. I decided to implement it using the smallest of the 5 models that I had found earlier, tinyllama-1.1b-chat-v1.0 Q6_K, to minimise the chances of having any memory issues.

You can find my prototype code here.

For the prototype, I decided to make very basic configurations for everything to develop it quickly, and I tested it by asking a question about Grace Hopper, one of the most influential programmers to have ever lived, in two versions of the prototype: one with Grace Hoppers Wikipedia page in its vector DB and one with Alan Turings Wikipedia page in its vector DB.

The question was:

What programming language was Grace Hopper known for creating?

As expected, the version with Alan Turings Wikipedia page failed miserably, hallucinating its entire response:

Grace Hopper was a computer scientist who is best known for her work on developing an early version of BASIC, which was one of the first low-level programming languages designed for mainframe computers. Hopper worked at Digital Equipment Corporation (DEC) in the mid-1960s and was responsible for creating several key BASIC versions that were used by DEC's mainframe computers. In 1973, she left DEC to work on a new low-level programming language called PASCAL, which became one of the most popular programming languages of its time. However, her contributions to BASIC are still remembered today as it was one of the first widely used and easy-to-understand programming languages for mainframe computers.

This behaviour of hallucinating will need to be fixed in the final product, likely through prompt engineering.

As expected, the version with Grace Hoppers Wikipedia page gave a much more accurate response but it unexpectedly contained far more spelling errors:

Grace Hopper is credited with creating the first computer programming languaige called FLOW-MATIC. This language was developed in 1954 and was based on English words. Hopper's work on this languaige led to a machine-indepenent programming language, COBOL, which used English words instead of symbols for input and output. Her compiler, A-0, also utilized the FLOW-MATIIC language as its basis for compiling programs into executable code. In 1952, Hopper published her first paper on the topic of compiler design, "The Compiler" in the IEEE Transactions on Computers, in which she discussed her work with A-0 and other programming languaiges.

Unfortunately, this response was not free from hallucinations either. Hoppers A-0 compiler did not use the FLOW-MATIC programming language, and she never published a paper called The Compiler in IEEE Transactions on Computers (a real journal that is not mentioned in the Wikipedia page). It looks like hallucinations are likely to be a major issue for this tool, but that is a problem I will solve when refining the AI.

On the bright side, inference was ~120 tokens/second, so at least this model will output words much faster than I can read them.

Conclusion

In this blog, I built a locally run prototype for my chatbot for querying unstructured text documents. It doesnt have a UI, and it hallucinates a lot, but it is nonetheless capable of querying unstructured text.

Its such a shame that after I had done the research and written all of the code, while I was writing this blog, I read about llmware, a very promising Python framework for building RAG pipelines with small models (sound familiar?). It was even chosen for GitHub Accelerator 2024, a competition for open-source projects on GitHub where chosen projects are given funding, mentorship, and access

to resources to help them grow their project. Since I had already built the prototype in LangChain, it didnt make much sense to tear it down and rebuild it in a fancy new framework that wasnt as tried-and-tested. Id love to try the framework out one day if I build another RAG application after this one.

In the next part of the STICI-note blog series, I will be building an evaluation suite to test and compare different inference models and vector DB configurations, so stay tuned and follow me on LinkedIn to be notified when it comes out!

Document 3 (Source: https://github.com/llmware-ai/llmware)

llmware

Building Enterprise RAG Pipelines with Small, Specialized Models
llmware provides a unified framework for building LLM-based applications (e.g, RAG, Agents), using small, specialized models that can be deployed privately, integrated with enterprise knowledge sources safely and securely, and cost-effectively tuned and adapted for any business process.

llmware has two main components:

RAG Pipeline - integrated components for the full lifecycle of connecting knowledge sources to generative AI models; and

50+ small, specialized models fine-tuned for key tasks in enterprise process automation, including fact-based question-answering, classification, summarization, and extraction.

By bringing together both of these components, along with integrating leading open source models and underlying technologies, llmware offers a comprehensive set of tools to rapidly build knowledge-based enterprise LLM applications.

Most of our examples can be run without a GPU server - get started right away on your laptop.

Join us on Discord | Watch Youtube Tutorials | Explore our Model Families on Huggingface

New to RAG? Check out the Fast Start video series

Multi-Model Agents with SLIM Models - Intro-Video

Intro to SLIM Function Call Models
Can't wait? Get SLIMs right away:

```
from llmware.models import ModelCatalog

ModelCatalog().get_llm_toolkit()  # get all SLIM models, delivered as small, fast quantized tools
ModelCatalog().tool_test_run("slim-sentiment-tool") # see the model in action with test script included
```

Key features
Writing code withllmware is based on a few main concepts:

Model Catalog: Access all models the same way with easy lookup, regardless of underlying implementation.
Library: ingest, organize and index a collection of knowledge at scale - Parse, Text Chunk and Embed.
Query: query libraries with mix of text, semantic, hybrid, metadata, and custom filters.
Prompt with Sources: the easiest way to combine knowledge retrieval with a LLM inference.
RAG-Optimized Models - 1-7B parameter models designed for RAG workflow integration and running locally.
Simple-to-Scale Database Options - integrated data stores from laptop to parallelized cluster.
Agents with Function Calls and SLIM Models
Start coding - Quick Start for RAG
What's New?
-Best New Small RAG Model - BLING finetune of Phi-3 - "bling-phi-3-gguf" - see the video

-Web Services with Agent Calls for Financial Research - end-to-end scenario - video and example

-Voice Transcription with WhisperCPP - getting_started, using_sample_files, and analysis_use_case with great_speeches_video

-Phi-3 GGUF Streaming Local Chatbot with UI - setup your own Phi-3-gguf chatbot on your laptop in minutes - example with video

-Small, specialized, function-calling Extract Model - introducing slim-extract - video and example

-LLM to Answer Yes/No questions - introducing slim-boolean model - video and example

-Natural Language Query to CSV End to End example - using slim-sql model - video and example and now using Custom Tables on Postgres example

-Multi-Model Agents with SLIM models - multi-step Agents with SLIMs on CPU - video - example

-OCR Embedded Document Images Example - systematically extract text from images embedded in documents example

-Enhanced Parser Functions for PDF, Word, Powerpoint and Excel - new text-chunking controls and strategies, extract tables, images, header text - example

-Agent Inference Server - set up multi-model Agents over Inference Server example

-GGUF - Getting Started - check out examples - GGUF (example) and Videos video

-Optimizing Accuracy of RAG Prompts - check out example and videos - part I and part II

Getting Started
Step 1 - Install llmware - pip3 install llmware or pip3 install 'llmware[full]'

note: starting with v0.3.0, we provide options for a core install (minimal set of dependencies) or full install (adds to the core with wider set of related python libraries).
Step 2- Go to Examples - Get Started Fast with 100+ 'Cut-and-Paste' Recipes
Step 3 - Tutorial Videos - check out our Youtube channel for high-impact 5-10 minute tutorials on the latest examples.
Working with the llmware Github repository

The llmware repo can be pulled locally to get access to all the examples, or to work directly with the latest version of the llmware code.

git clone git@github.com:llmware-ai/llmware.git
We have provided a welcome_to_llmware automation script in the root of the repository folder. After cloning:

On Windows command line: .\welcome_to_llmware_windows.sh
On Mac / Linux command line: sh ./welcome_to_llmware.sh
Alternatively, if you prefer to complete setup without the welcome automation script, then the next steps include:

install requirements.txt - inside the /llmware path - e.g., pip3 install -r llmware/requirements.txt

install requirements_extras.txt - inside the /llmware path - e.g., pip3 install -r llmware/requirements_extras.txt (Depending upon your use case, you may not need all or any of these installs, but some of these will be used in the examples.)

run examples - copy one or more of the example .py files into the root project path. (We have seen several IDEs that will attempt to run interactively from the nested /example path, and then not have access to the /llmware module - the easy fix is to just copy the example you want to run into the root path).

install vector db - no-install vector db options include milvus lite, chromadb, faiss and lancedb - which do not require a server install, but do require that you install the python sdk library for that vector db, e.g., pip3 install pymilvus, or pip3 install chromadb. If you look in examples/Embedding, you will see examples for getting started with various vector DB, and in the root of the repo, you will see easy-to-get-started docker compose scripts for installing milvus, postgres/pgvector, mongo, qdrant, neo4j, and redis.

Note: we have seen recently issues with Pytorch==2.3 on some platforms - if you run into any issues, we have seen that uninstalling Pytorch and downleveling to Pytorch==2.1 usually solves the problem.

Data Store Options

Fast Start: use SQLite3 and ChromaDB (File-based) out-of-the-box - no install required

Speed + Scale: use MongoDB (text collection) and Milvus (vector db) - install with Docker Compose

Postgres: use Postgres for both text collection and vector DB - install with Docker Compose

Mix-and-Match: LLMWare supports 3 text collection databases (Mongo, Postgres, SQLite) and 10 vector databases (Milvus, PGVector-Postgres, Neo4j, Redis, Mongo-Atlas, Qdrant, Faiss, LanceDB, ChromaDB and Pinecone)

Meet our Models

SLIM model series: small, specialized models fine-tuned for function calling and multi-step, multi-model Agent workflows.

DRAGON model series: Production-grade RAG-optimized 6-7B parameter models - "Delivering RAG on ..." the leading foundation base models.

BLING model series: Small CPU-based RAG-optimized, instruct-following 1B-3B parameter models.

Industry BERT models: out-of-the-box custom trained sentence transformer embedding models fine-tuned for the following industries: Insurance, Contracts, Asset Management, SEC.

GGUF Quantization: we provide 'gguf' and 'tool' versions of many SLIM, DRAGON and BLING models, optimized for CPU deployment.

Using LLMs and setting-up API keys & secrets

LLMWare is an open platform and supports a wide range of open source and proprietary models. To use LLMWare, you do not need to use any proprietary LLM - we would encourage you to experiment with SLIM, BLING, DRAGON, Industry-BERT, the GGUF examples, along with bringing in your favorite models from HuggingFace and Sentence Transformers.

If you would like to use a proprietary model, you will need to provide your own API Keys. API keys and secrets for models, aws, and pinecone can be set-up for use in environment variables or passed directly to method calls.

Roadmap - Where are we going ...

Interested in contributing to llmware? Information on ways to participate can be found in our Contributors Guide. As with all aspects of this project, contributing is governed by our Code of Conduct.

Questions and discussions are welcome in our github discussions.

Release notes and Change Log

See also additional deployment/install release notes in wheel_archives

Thursday, June 6 - v0.3.1-WIP

Added module 3 to Fast Start example series examples 7-9 on Agents & Function Calls

Added reranker Jina model for in-memory semantic similarity RAG - see example

Changes merged into main branch - expected next pypi release at end of week

Tuesday, June 4 - v0.3.0

Added support for new Milvus Lite embedded 'no-install' database - see example.

Added two new SLIM models to catalog and agent processes - 'q-gen' and 'qa-gen'

Updated model class instantiation to provide more extensibility to add new classes in different modules

New welcome_to_llmware.sh and welcome_to_llmware_windows.sh fast install scripts

Enhanced Model class base with new configurable post_init and register methods

Created InferenceHistory to track global state of all inferences completed

Multiple improvements and updates to logging at module level

Note: starting with v0.3.0, pip install provides two options - a base minimal install pip3 install llmware which will support most use cases, and a larger install pip3 install 'llmware[full]' with other commonly-used libraries.

Wednesday, May 22 - v0.2.15

Improvements in Model class handling of Pytorch and Transformers dependencies (just-in-time loading, if needed)

Expanding API endpoint options and inference server functionality - see new client access options and server_launch

Saturday, May 18 - v0.2.14

New OCR image parsing methods with example

Adding first part of logging improvements (WIP) in Configs and Models.

New embedding model added to catalog - industry-bert-loans.

Updates to model import methods and configurations.

Sunday, May 12 - v0.2.13

New GGUF streaming method with basic example and phi3 local chatbot

Significant cleanups in ancillary imports and dependencies to reduce install complexity - note: the updated requirements.txt and setup.py files.

Defensive code to provide informative warning of any missing dependencies in specialized parts of the code, e.g., OCR, Web Parser.

Updates of tests, notice and documentation.

OpenAIConfigs created to support Azure OpenAI.

Sunday, May 5 - v0.2.12 Update

Launched "bling-phi-3" and "bling-phi-3-gguf" in ModelCatalog - newest and most accurate BLING/DRAGON model

New long document summarization method using slim-summary-tool example

New Office (Powerpoint, Word, Excel) sample files example

Added support for Python 3.12

Deprecated faiss and replaced with 'no-install' chromadb in Fast Start examples

Refactored Datasets, Graph and Web Services classes

Updated Voice parsing with WhisperCPP into Library

Monday, April 29 - v0.2.11 Update

Updates to gguf libs for Phi-3 and Llama-3

Added Phi-3 example and Llama-3 example and Quantized Versions to Model Catalog

Integrated WhisperCPP Model class and prebuilt shared libraries - getting-started-example

New voice sample files for testing - example

Improved CUDA detection on Windows and safety checks for older Mac OS versions

Monday, April 22 - v0.2.10 Update

Updates to Agent class to support Natural Language queries of Custom Tables on Postgres example

New Agent API endpoint implemented with LLMWare Inference Server and new Agent capabilities

example

Tuesday, April 16 - v0.2.9 Update

New CustomTable class to rapidly create custom DB tables in conjunction with LLM-based workflows.

Enhanced methods for converting CSV and JSON/JSONL files into DB tables.

See new examples Creating Custom Table example

Tuesday, April 9 - v0.2.8 Update

Office Parser (Word Docx, Powerpoint PPTX, and Excel XLSX) - multiple improvements - new libs + Python method.

Includes: several fixes, improved text chunking controls, header text extraction and configuration options.

Generally, new office parser options conform with the new PDF parser options.

Please see Office Parsing Configs example

Wednesday, April 3 - v0.2.7 Update

PDF Parser - multiple improvements - new libs + Python methods.

Includes: UTF-8 encoding for European languages.

Includes: Better text chunking controls, header text extraction and configuration options.

Please see PDF Parsing Configs example for more details.

Note: deprecating support for aarch64-linux (will use 0.2.6 parsers). Full support going forward for Linux Ubuntu20+ on x86_64 + with CUDA.

Friday, March 22 - v0.2.6 Update

New SLIM models: summary, extract, xsum, boolean, tags-3b, and combo sentiment-ner.

New logit and sampling analytics.

New SLIM examples showing how to use the new models.

Thursday, March 14 - v0.2.5 Update

Improved support for GGUF on CUDA (Windows and Linux), with new prebuilt binaries and exception handling.

Enhanced model configuration options (sampling, temperature, top logit capture).

Added full back-level support for Ubuntu 20+ with parsers and GGUF engine.

Support for new Anthropic Claude 3 models.

New retrieval methods: document_lookup and aggregate_text.

New model: bling-stablelm-3b-tool - fast, accurate 3b quantized question-answering model - one of our new favorites.

Wednesday, February 28 - v0.2.4 Update

Major upgrade of GGUF Generative Model class - support for Stable-LM-3B, CUDA build options, and better control over sampling strategies.

Note: new GGUF llama.cpp built libs packaged with build starting in v0.2.4.

Improved GPU support for HF Embedding Models.

Friday, February 16 - v0.2.3 Update

Added 10+ embedding models to ModelCatalog - nomic, jina, bge, gte, ember and uae-large.

Updated OpenAI support >=1.0 and new text-3 embedding models.

SLIM model keys and output_values now accessible in ModelCatalog.

Updating encodings to 'utf-8-sig' to better handle txt/csv files with bom.

Supported Operating Systems: MacOS (Metal and x86), Linux (x86 and aarch64), Windows

note on Linux: we test most extensively on Ubuntu 22 and now Ubuntu 20 and recommend where possible

if you need another Linux version, please raise an issue - we will prioritize testing and ensure support.

Supported Vector Databases: Milvus, Postgres (PGVector), Neo4j, Redis, LanceDB, ChromaDB, Qdrant, FAISS, Pinecone, Mongo Atlas Vector Search

Supported Text Index Databases: MongoDB, Postgres, SQLite

Optional

Docker

To enable the OCR parsing capabilities, install Tesseract v5.3.3 and Poppler v23.10.0 native packages.

Change Log

Latest Updates - 19 Jan 2024 - llmware v0.2.0

Added new database integration options - Postgres and SQlite

Improved status update and parser event logging options for parallelized parsing

Significant enhancements to interactions between Embedding + Text collection databases

Improved error exception handling in loading dynamic modules

Latest Updates - 15 Jan 2024: llmware v0.1.15

Enhancements to dual pass retrieval queries

Expanded configuration objects and options for endpoint resources

Latest Updates - 30 Dec 2023: llmware v0.1.14

Added support for Open Chat inference servers (compatible with OpenAI API)

Improved capabilities for multiple embedding models and vector DB configurations

Added docker-compose install scripts for PGVector and Redis vector databases

Added 'bling-tiny-llama' to model catalog

Latest Updates - 22 Dec 2023: llmware v0.1.13

Added 3 new vector databases - Postgres (PG Vector), Redis, and Qdrant

Improved support for integrating sentence transformers directly in the model catalog

Improvements in the model catalog attributes

Multiple new Examples in Models & Embeddings, including GGUF, Vector database, and model catalog

17 Dec 2023: llmware v0.1.12

dragon-deci-7b added to catalog - RAG-finetuned model on high-performance new 7B model base from Deci

New GGUFGenerativeModel class for easy integration of GGUF Models

Adding prebuilt llama_cpp / ctransformer shared libraries for Mac M1, Mac x86, Linux x86 and Windows

3 DRAGON models packaged as Q4_K_M GGUF models for CPU laptop use (dragon-mistral-7b, dragon-llama-7b, dragon-yi-6b)

4 leading open source chat models added to default catalog with Q4_K_M

8 Dec 2023: llmware v0.1.11

New fast start examples for high volume Document Ingestion and Embeddings with Milvus.

New LLMWare 'Pop up' Inference Server model class and example script.

New Invoice Processing example for RAG.

Improved Windows stack management to support parsing larger documents.

Enhancing debugging log output mode options for PDF and Office parsers.

30 Nov 2023: llmware v0.1.10

Windows added as a supported operating system.

Further enhancements to native code for stack management.

Minor defect fixes.

24 Nov 2023: llmware v0.1.9

Markdown (.md) files are now parsed and treated as text files.

PDF and Office parser stack optimizations which should avoid the need to set ulimit -s.

New llmware_models_fast_start.py example that allows discovery and selection of all llmware HuggingFace models.

Native dependencies (shared libraries and dependencies) now included in repo to faciliate local development.

Updates to the Status class to support PDF and Office document parsing status updates.

Minor defect fixes including image block handling in library exports.

17 Nov 2023: llmware v0.1.8

Enhanced generation performance by allowing each model to specific the trailing space parameter.

Improved handling for eos_token_id for llama2 and mistral.

Improved support for Hugging Face dynamic loading

New examples with the new llmware DRAGON models.

14 Nov 2023: llmware v0.1.7

Moved to Python Wheel package format for PyPi distribution to provide seamless installation of native dependencies on all supported platforms.

ModelCatalog enhancements:

OpenAI update to include newly announced turbo 4 and 3.5 models.

Cohere embedding v3 update to include new Cohere embedding models.

BLING models as out-of-the-box registered options in the catalog. They can be instantiated like any other model, even without the hf=True flag.

Ability to register new model names, within existing model classes, with the register method in ModelCatalog.

Prompt enhancements:

evidence_metadata added to prompt_main output dictionaries allowing prompt_main responses to be plug into the evidence and fact-checking steps without modification.

API key can now be passed directly in a prompt.load_model(model_name, api_key = [my-api-key])

LLMWareInference Server - Initial delivery:

New Class for LLMWareModel which is a wrapper on a custom HF-style API-based model.

LLMWareInferenceServer is a new class that can be instantiated on a remote (GPU) server to create a testing API-server that can be integrated into any Prompt workflow.

03 Nov 2023: llmware v0.1.6

Updated packaging to require mongo-c-driver 1.24.4 to temporarily workaround segmentation fault with mongo-c-driver 1.25.

Updates in python code needed in anticipation of future Windows support.

27 Oct 2023: llmware v0.1.5

Four new example scripts focused on RAG workflows with small, fine-tuned instruct models that run on a laptop (llmware BLING models).

Expanded options for setting temperature inside a prompt class.

Improvement in post processing of Hugging Face model generation.

Streamlined loading of Hugging Face generative models into prompts.

Initial delivery of a central status class: read/write of embedding status with a consistent interface for

callers.

Enhanced in-memory dictionary search support for multi-key queries.

Removed trailing space in human-bot wrapping to improve generation quality in some fine-tuned models.

Minor defect fixes, updated test scripts, and version update for Werkzeug to address dependency security alert.

20 Oct 2023: llmware v0.1.4

GPU support for Hugging Face models.

Defect fixes and additional test scripts.

13 Oct 2023: llmware v0.1.3

MongoDB Atlas Vector Search support.

Support for authentication using a MongoDB connection string.

Document summarization methods.

Improvements in capturing the model context window automatically and passing changes in the expected output length.

Dataset card and description with lookup by name.

Processing time added to model inference usage dictionary.

Additional test scripts, examples, and defect fixes.

06 Oct 2023: llmware v0.1.1

Added test scripts to the github repository for regression testing.

Minor defect fixes and version update of Pillow to address dependency security alert.

02 Oct 2023: llmware v0.1.0 Initial release of llmware to open source!!

Document 4 (Source: https://docs.marimo.io/recipes.html)

Recipes

This page includes code snippets or recipes for a variety of common tasks. Use them as building blocks or examples when making your own notebooks.

In these recipes, each code block represents a cell.

Control Flow

Show an output conditionally

Use cases. Hide an output until a condition is met (e.g., until algorithm parameters are valid), or show different outputs depending on the value of a UI element or some other Python object

Recipe.

Use an if expression to choose which output to show.

```
# condition is a boolean, True of False
condition = True
"condition is True" if condition else None
```

Run a cell on a timer

Use cases.

Load new data periodically, and show updated plots or other outputs. For example, in a dashboard monitoring a training run, experiment trial, real-time weather data,

Run a job periodically

Recipe.

Import packages

```
import marimo as mo
```

Create a mo.ui.refresh timer that fires once a second:

```
refresh = mo.ui.refresh(default_interval="1s")
# This outputs a timer that fires once a second
refresh
```

Reference the timer by name to make this cell run once a second

```
import random
```

This cell will run once a second!
refresh

mo.md("#" + "" * random.randint(1, 10))
Require form submission before sending UI value
Use cases. UI elements automatically send their values to the Python when they are interacted with, and run all cells referencing the elements. This makes marimo notebooks responsive, but it can be an issue when the downstream cells are expensive, or when the input (such as a text box) needs to be filled out completely before it is considered valid. Forms let you gate submission of UI element values on manual confirmation, via a button press.

Recipe.

Import packages

import marimo as mo
Create a submittable form.

form = mo.ui.text(label="Your name").form()
form
Get the value of the form.

form.value
Stop execution of a cell and its descendants
Use cases. For example, dont run a cell or its descendants if a form is unsubmitted.

Recipe.

Import packages

import marimo as mo
Create a submittable form.

form = mo.ui.text(label="Your name").form()

form

Use mo.stop to stop execution when the form is unsubmitted.

mo.stop(form.value is None, mo.md("Submit the form to continue"))

mo.md(f"Hello, {form.value}!")

Grouping UI elements together

Create an array of UI elements

Use cases. In order to synchronize UI elements between the frontend and backend (Python), marimo requires you to assign UI elements to global variables. But sometimes you dont know the number of elements to make until runtime: for example, maybe you want o make a list of sliders, and the number of sliders to make depends on the value of some other UI element.

You might be tempted to create a Python list of UI elements, such as l = [mo.ui.slider(1, 10) for i in range(number.value)]: however, this wont work, because the sliders are not bound to global variables.

For such cases, marimo provides the higher-order UI element mo.ui.array, which lets you make a new UI element out of a list of UI elements: l = mo.ui.array([mo.ui.slider(1, 10) for i in range(number.value)]). The value of an array element is a list of the values of the elements it wraps (in this case, a list of the slider values). Any time you interact with any of the UI elements in the array, all cells referencing the array by name (in this case, l) will run automatically.

Recipe.

Import packages.

import marimo as mo

Use mo.ui.array to group together many UI elements into a list.

import random

```
# instead of random.randint, in your notebook you'd use the value of
# an upstream UI element or other Python object
array = mo.ui.array([mo.ui.text() for i in range(random.randint(1, 10))])
array
```

Get the value of the UI elements using array.value

```
array.value
```

Create a dictionary of UI elements

Use cases. Same as for creating an array of UI elements, but lets you name each of the wrapped elements with a string key.

Recipe.

Import packages.

```
import marimo as mo
```

Use mo.ui.dictionary to group together many UI elements into a list.

```
import random

# instead of random.randint, in your notebook you'd use the value of
# an upstream UI element or other Python object
dictionary = mo.ui.dictionary({str(i): mo.ui.text() for i in range(random.randint(1, 10))})
dictionary
```

Get the value of the UI elements using dictionary.value

```
dictionary.value
```

Embed a dynamic number of UI elements in another output

Use cases. When you want to embed a dynamic number of UI elements in other outputs (like tables or markdown).

Recipe.

Import packages

```
import marimo as mo
```

Group the elements with mo.ui.dictionary or mo.ui.array, then retrieve them from the container and display them elsewhere.

```
import random

n_items = random.randint(2, 5)

# Create a dynamic number of elements using `mo.ui.dictionary` and
# `mo.ui.array`
elements = mo.ui.dictionary(
    {
        "checkboxes": mo.ui.array([mo.ui.checkbox() for _ in range(n_items)]),
        "texts": mo.ui.array(
            [mo.ui.text(placeholder="task ...") for _ in range(n_items)]
        ),
    }
)

mo.md(
    f"""
    Here's a TODO list of {n_items} items\n\n
    """
    + "\n\n".join(
        # Iterate over the elements and embed them in markdown
        [
            f"{checkbox} {text}"
            for checkbox, text in zip(
                elements["checkboxes"], elements["texts"]
            )
        ]
```

```
    )
)
```

Get the value of the elements

```
elements.value
```

Create a hstack (or vstack) of UI elements with on_change handlers

Use cases. Arrange a dynamic number of UI elements in a hstack or vstack, for example some number of buttons, and execute some side-effect when an element is interacted with, e.g. when a button is clicked.

Recipe.

Import packages

```
import marimo as mo
```

Create buttons in mo.ui.array and pass them to hstack a regular Python list wont work. Make sure to assign the array to a global variable.

```
import random



# Create a state object that will store the index of the
# clicked button
get_state, set_state = mo.state(None)

# Create an mo.ui.array of buttons - a regular Python list won't work.
buttons = mo.ui.array(
    [
        mo.ui.button(
            label="button " + str(i), on_change=lambda v, i=i: set_state(i)
        )
        for i in range(random.randint(2, 5))
    ]
```

)

mo.hstack(buttons)
Get the state value

get_state()
Create a table column of buttons with on_change handlers
Use cases. Arrange a dynamic number of UI elements in a column of a table, and execute some side-effect when an element is interacted with, e.g. when a button is clicked.

Recipe.

Import packages

import marimo as mo
Create buttons in mo.ui.array and pass them to mo.ui.table. Make sure to assign the table and array to global variables

import random

```
# Create a state object that will store the index of the
# clicked button
get_state, set_state = mo.state(None)

# Create an mo.ui.array of buttons - a regular Python list won't work.
buttons = mo.ui.array(
    [
        mo.ui.button(
            label="button " + str(i), on_change=lambda v, i=i: set_state(i)
        )
        for i in range(random.randint(2, 5))
    ]
```

```
)

# Put the buttons array into the table
table = mo.ui.table(
    {
        "Action": ["Action Name"] * len(buttons),
        "Trigger": list(buttons),
    }
)
table
```

Get the state value

```
get_state()
```

Create a form with multiple UI elements

Use cases. Combine multiple UI elements into a form so that submission of the form sends all its elements to Python.

Recipe.

Import packages.

```
import marimo as mo
```

Use mo.ui.form and Html.batch to create a form with multiple elements.

```
form = mo.md(
    r"""
    Choose your algorithm parameters:

    - $\epsilon$: {epsilon}
    - $\delta$: {delta}
    """
).batch(epsilon=mo.ui.slider(0.1, 1, step=0.1), delta=mo.ui.number(1, 10)).form()
form
```

Get the submitted form value.

form.value

Working with buttons

Create a button that triggers computation when clicked

Use cases. To trigger a computation on button click and only on button click, use mo.ui.run_button().

Recipe.

Import packages

```
import marimo as mo
```

Create a run button

```
button = mo.ui.run_button()
button
```

Run something only if the button has been clicked.

```
mo.stop(not button.value, "Click 'run' to generate a random number")

import random
random.randint(0, 1000)
```

Create a counter button

Use cases. A counter button, i.e. a button that counts the number of times it has been clicked, is a helpful building block for reacting to button clicks (see other recipes in this section).

Recipe.

Import packages

```
import marimo as mo
```

Use mo.ui.button and its on_click argument to create a counter button.

```
# Initialize the button value to 0, increment it on every click
button = mo.ui.button(value=0, on_click=lambda count: count + 1)
button
```

Get the button value

```
button.value
```

Create a toggle button

Use cases. Toggle between two states using a button with a button that toggles between True and False. (Tip: you can also just use mo.ui.switch.)

Recipe.

Import packages

```
import marimo as mo
```

Use mo.ui.button and its on_click argument to create a toggle button.

```
# Initialize the button value to False, flip its value on every click.
button = mo.ui.button(value=False, on_click=lambda value: not value)
button
```

Toggle between two outputs using the button value.

```
mo.md("True!") if button.value else mo.md("False!")
```

Re-run a cell when a button is pressed

Use cases. For example, you have a cell showing a random sample of data, and you want to resample on button press.

Recipe.

Import packages

```
import marimo as mo
```

Create a button without a value, to function as a trigger.

button = mo.ui.button()

button

Reference the button in another cell.

the button acts as a trigger: every time it is clicked, this cell is run

button

Replace with your custom lgic

import random

random.randint(0, 100)

Run a cell when a button is pressed, but not before

Use cases. Wait for confirmation before executing downstream cells (similar to a form).

Recipe.

Import packages

import marimo as mo

Create a counter button.

button = mo.ui.button(value=0, on_click=lambda count: count + 1)

button

Only execute when the count is greater than 0.

Don't run this cell if the button hasn't been clicked, using mo.stop.

Alternatively, use an if expression.

mo.stop(button.value == 0)

mo.md(f"The button was clicked {button.value} times")

Reveal an output when a button is pressed

Use cases. Incrementally reveal a user interface.

Recipe.

Import packages

import marimo as mo
Create a counter button.

button = mo.ui.button(value=0, on_click=lambda count: count + 1)
button
Show an output after the button is clicked.

mo.md("#" + "" * button.value) if button.value > 0 else None
Caching
Cache expensive computations
Use case. Because marimo runs cells automatically as code and UI elements change, it can be helpful to cache expensive intermediate computations. For example, perhaps your notebook computes t-SNE, UMAP, or PyMDE embeddings, and exposes their parameters as UI elements. Caching the embeddings for different configurations of the elements would greatly speed up your notebook.

Recipe.

Use functools to cache function outputs given inputs.

import functools

@functools.cache
def compute_predictions(problem_parameters):
  # replace with your own function/parameters
  ...
Whenever compute_predictions is called with a value of problem_parameters it has not seen, it will compute the predictions and store them in a cache. The next time it is called with the same parameters, instead of recomputing the predictions, it will return the previously computed value from

the cache.

See our best practices guide to learn more.

Document 5 (Source: https://towardsdatascience.com/how-to-maximize-your-impact-as-a-data-scientist-3881995a9cb1)

How to Maximize Your Impact as a Data Scientist

One of the hardest pills to swallow as an Individual Contributor (IC) at work is that nobody cares about the hard work you put in. They dont even care about your output; they care about the impact you drive.

Whats the difference? Your output is the analysis you deliver, or the lines of code you write. Your impact is the decision your analysis helps the CEO make, or the revenue the new product feature is generating.

Image by author

If you want to establish yourself as a high performer and accelerate your career as a Data Scientist, its key to focus on impact.

In this post Ill go over the following:

Why prioritizing impact matters not just for managers, but also ICs

Why focusing on impact is hard

How to maximize your impact

How to overcome common challenges in driving real impact

Lets dive in.

Get an email whenever Torsten Walbaum publishes.

Get an email whenever Torsten Walbaum publishes. By signing up, you will create a Medium account if you don't already

medium.com

Why should I focus on impact; isnt that my managers job?

Of course you can leave it to your manager to worry about impact. But stepping up comes with some real benefits for your career:

Reduced frustration & burn-out: Putting a lot of work into a project and then feeling like it didnt move the needle is one of the most frustrating feelings in any job.

Promotions: Promotions are heavily tied to impact. And if you want to become a manager, youll need to show that you understand what drives business outcomes and can allocate resources accordingly.

Internal opportunities: People around you notice if you are having an outsized impact, and youll increase your chances of receiving internal offers. My promotion to Director happened because the CMO noticed my work on the BizOps team and asked me to move into the Marketing org to build out a Strategy & Analytics team.

External opportunities: Prospective employers dont focus on what responsibilities you had, but what your impact was. After all, they are trying to figure out how you can help their business.

Why isnt everyone doing this?

Because its hard.

We are used to thinking about inputs and outputs rather than impact in our daily lives (I went to the gym or I did three loads of laundry) and we carry that mindset over to our jobs.

More importantly, it gives us a sense of control. Its fully under your control to work hard on the project, and maybe to create the final deliverable, but you cant guarantee that it will actually move the business forward.

It can also feel like were doing someone elses job. You built the dashboard; now its the other teams problem how theyre going use it and get value from it. You can definitely take this stance; but dont you want to see your work move the needle?

Lastly, sometimes its unclear what impact even looks like for our role because we feel too disconnected from the business outcomes; Ill get into this below.

How can I become more impact-focused?

Step 1: Understand what impact looks like for your role and measure your success accordingly

Stop thinking about productivity metrics like I launched 5 experiments or I built this model and hold yourself accountable to driving impact.

But what does that look like for a Data Scientist? For other roles its easy; Account Executives have sales quotas and Growth Marketing Managers have lead generation targets.

But Data Science, at its core, is a function that supports other teams. As a result, there are two levels of impact:

Image by author

Did your work change anything for the better for your business partners? E.g.:

Did your analysis change the roll-out strategy of the new product?

Did your model improve forecast accuracy?

Does your dashboard save the team hours every week that they used to spend on manual data pulls?

Did your work help move the needle on downstream business metrics? E.g.:

Youre a Marketing Data Scientist? Assume youre on the hook for hitting lead and opportunity targets, and improving Marketing efficiency

Youre doing Analytics for the Customer Support org? Start obsessing about response times and satisfaction scores.

You dont have to be solely responsible for something in order to take (partial) credit for it. If you provided the analysis that resulted in a pricing change that saved the company millions, then you deserve part of the credit for that impact.

You might not feel the consequences of missing these downstream targets as immediately as your stakeholders, but since your long-term career trajectory is still tied to driving impact, it helps to adopt this outcome-focused mindset.

Once you start doing this, youll notice more inefficiencies you can help address, or new opportunities for growth.

Step 2: Ensure your work solves a real business problem

Youll likely know this situation: Instead of approaching you with a problem, people ask you for a specific deliverable. An analysis, a model, a dashboard.

If you blindly execute what they ask, you might realize too late that it wont lead to tangible business impact. Maybe the problem they are trying to solve is not that important in the grand scheme of things, or there is a better way to approach it.

So what can you do?

Act like an owner. Understand the actual problem behind the request, and ask yourself what business priority this supports.

If you are early in your career then your manager should ideally help with this. But dont rely on this: Managers dont always do a perfect job, and youll be the one to feel the consequences of badly scoped work.

This requires you to understand company level priorities and the priorities of other orgs and teams. Take notes during All Hands meetings etc. to understand the big picture, and get your hands on other teams planning materials to get an idea of what theyre trying to accomplish in the next 12 quarters.

Step 3: Ensure there is buy-in for your work

Even if your work directly supports company-level priorities, youll be in for a bad time if key stakeholders are not bought in.

You dont want to be in a situation where you finish the work and then realize that another team is blocking the implementation because they have concerns you didnt address. To avoid this, youll:

Need to understand whose support you need, and

Get them onboard from the get-go

This is a complex topic in itself; Ill write a separate deep dive on how to drive alignment and get buy-in from other teams in the near future.

Step 4: Focus your time on the highest-impact thing

No matter what role youre in, youre likely juggling multiple priorities. To maximize your impact, you need to ensure you spend the majority of your time on the most important thing.

As with many things, this is easier said than done though, so lets talk about what that looks like concretely.

Ad-hoc requests vs. strategic work

Its easy to get caught up in the craziness of daily business only to realize you didnt make any progress on the big, strategic project you actually care about.

This is all too common; none of us get to sit in our ivory tower and chip away at our projects undisturbed. Plus, ad-hoc work is impactful, too; while its less exciting than strategic projects, its what keeps the business running.

Still, if you find yourself spending the majority of your time fielding these ad-hoc issues, its time to talk to your manager. Im sure your manager would rather help protect your bandwidth than have you 1) miss your deadlines on your key projects and 2) quit eventually from frustration.

Image by author

Dont cry over spilled milk

Another common challenge comes from the sunk cost fallacy. You invested a lot of time into a project, but it doesnt seem to be going anywhere. Maybe you realized the premise didnt make as much sense as you thought, or the priorities of the business have changed since you started the work.

Instead of talking to your manager and stakeholders about changing the scope of the project or abandoning it altogether, youre doubling down to get it over the finish line. After all, you dont want all of your effort to go to waste. Sound familiar?

Economists (and Poker players) figured out a long time ago that this is a dangerous trap. When prioritizing your time, ignore how much effort your already put in and focus on where the next hour of work will yield the most impact.

Things to watch out for (impact killers)

How do you minimize the odds of wasting time on a project that wont lead to impact? There are a few warning signs:

Academic projects: Any time a project is pitched to you along the lines of This would be interesting to understand you should be careful; projects that purely improve the understanding of an issue without tying it back to the business are a waste of time and source of frustration in my experience

Overly ambitious project scope: At Uber, everyone always wanted to understand what the best driver incentive type is. Many people worked on this over the years, but it never led anywhere. There was no simple one-size-fits-all answer to this question, and the projects that led to actual impact were much more concrete, tactical optimizations

The customer or deliverable are not defined: If its not clear who the end user of your work is (are you doing this for your manager, leadership, or another team?), or youre unsure what exactly youre supposed to deliver, it should raise a red flag. This is typically a sign that the project needs more scoping work before someone should start running with it

Common Challenges and How to Address Them

We talked about general frameworks to maximize impact. But how do you make actual, specific projects more impactful?

Many times, projects fail close to the finish line. Impact doesnt materialize automatically, so you need to put in the final bit of work to ensure your work gets adopted. Doing this has an extremely high return on the time you invest since you already did the hard work to produce the deliverable and only need to close the loop with stakeholders.

Image by author

To make things more tangible, I am going to go through a few types of common deliverables, touch on where they typically fail to create impact and propose what you can do about it:

1. You create a comprehensive analysis but nobody is acting on it

Problem: This is common with analyses that dont have a clear recommendation. If you simply outline the data and potential paths forward, you are expecting your audience to do all of the heavy lifting.

Solution: Your work starts adding real value for them once you take that work off their plate. Always give a clear recommendation; you can caveat it and show alternatives in the appendix, but you need to take a stance.

2. You ran an experiment but nobody is using the results

Problem: Many experiments conclude with a metrics read-out by Data Science. More often than not, this is a metrics dump with a lot of information, but little interpretation or context.

Solution: Help your business partners interpret the results, and tell them how it affects what they care about.

How should they think about the statistical significance or lack thereof?

Is the observed lift good compared to other changes you tested and shipped?

What is your recommendation for next steps? What does the experiment result mean for this person or team specifically?

Remember, you are the subject matter expert and shouldnt expect non-analytical audiences to interpret raw experiment data. Telling your stakeholders what the result means for them will increase chances they will act on it.

3. You built a predictive model, but the team you built it for is not using it

Problem: When predictive models dont get used, its often because of a lack of trust in the model output.

ML models themselves tend to be black boxes, and if teams dont understand how the outputs were generated and whether they are reliable, they are hesitant to rely on them. Even if your model is not using ML and lives in a spreadsheet: If people dont know how it works, theyll be suspicious.

Solution: Its all about involving stakeholders in the process and building trust.

Involve stakeholders in the model development from the get-go to get them comfortable and address any concerns early on

Demystify the output; for example, you can extract the top model features and explain them

Sanity-check predictions and compare them to intuition. For example, if you forecast sales but your model predicts a different seasonality pattern from previous years, youll need to be able to explain why, or youll lose trust. In my experience, this is more impactful than just sharing performance metrics like the accuracy of the model

Having a structured playbook for how to do this will make your life easier, so Ill cover this in a separate post in the near future.

4. You created a dashboard but nobody is looking at it

Problem: If a dashboard doesnt get used, its likely one of these things is true:

The dashboard doesnt directly address an urgent business use case

You didnt involve your stakeholders along the way (e.g. by sharing mock-ups and drafts for feedback) and the final product is not what they were hoping for

The dashboard is complex and your users dont understand how to get what they need

Solution: To address #1 and #2, start with user research to understand pain points and potential use cases of the dashboard, and involve your stakeholders during development.

With regards to #3, a simpler dashboard that users are comfortable with beats a more advanced one that doesnt get used. If you cannot (or dont want to) simplify the dash further, youll need to train your users on the functionality and shadow them to understand any points of friction.

A dashboard is not done when you ship it for the first time, but needs to be improved over time based on users needs and feedback.

Closing Thoughts

Focusing on impact is scary since we leave the world of controllable inputs behind, but its what ultimately gets you promotions and new job opportunities.

And isnt it nice when your work actually feels like it moves the needle?

For more hands-on analytics advice, consider following me here on Medium, on LinkedIn or on Substack.

Document 6 (Source: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683)

Why do we need to regulate the use of Artificial Intelligence?

The potential benefits of Artificial Intelligence (AI) for our societies are manifold from improved medical care to better education. Faced with the rapid technological development of AI, the EU decided to act as one to harness these opportunities.

The EU AI Act is the world's first comprehensive AI law. It aims to address risks to health, safety and fundamental rights. The regulation also protects democracy, rule of law and the environment.

While most AI systems will pose low to no risk, certain AI systems create risks that need to be addressed to avoid undesirable outcomes.

For example, the opacity of many algorithms may create uncertainty and hamper the effective enforcement of the existing legislation on safety and fundamental rights. Responding to these challenges, legislative action was needed to ensure a well-functioning internal market for AI systems where both benefits and risks are adequately addressed.

This includes applications such as biometric identification systems or AI decisions touching on important personal interests, such as in the areas of recruitment, education, healthcare, or law enforcement.

Recent advancements in AI gave rise to ever more powerful Generative AI. So-called general-purpose AI models that are being integrated in numerous AI systems are becoming too important for the economy and society not to be regulated. In light of potential systemic risks, the EU puts in place effective rules and oversight.

Which risks will the new AI rules address?

The uptake of AI systems has a strong potential to bring societal benefits, economic growth and enhance EU innovation and global competitiveness. However, in certain cases, the specific characteristics of certain AI systems may create new risks related to user safety and fundamental rights. Some powerful AI models that are being widely used could even pose systemic risks.

This leads to legal uncertainty for companies and potentially slower uptake of AI technologies by businesses and citizens, due to the lack of trust. Disparate regulatory responses by national authorities would risk fragmenting the internal market.

To whom does the AI Act apply?
The legal framework will apply to both public and private actors inside and outside the EU as long as the AI system is placed on the Union market or its use affects people located in the EU.

It can concern both providers (e.g. a developer of a CV-screening tool) and deployers of high-risk AI systems (e.g. a bank buying this screening toolImporters of AI systems will also have to ensure that the foreign provider has already carried out the appropriate conformity assessment procedure, bears a European Conformity (CE) marking and is accompanied by the required documentation and instructions of use.

In addition, certain obligations are foreseen for providers of general-purpose AI models, including large generative AI models.

Providers of free and open-source models are exempted from most of these obligations. This exemption does not cover obligations for providers of general purpose AI models with systemic risks.

Obligations also do not apply to research, development and prototyping activities preceding the release on the market, and the regulation furthermore does not apply to AI systems that are exclusively for military, defence or national security purposes, regardless of the type of entity carrying out those activities.

What are the risk categories?
The Commission proposes a riskbased approach, with four levels of risk for AI systems, as well as

an identification of risks specific to general purpose models:

Minimal risk: All other AI systems can be developed and used subject to the existing legislation without additional legal obligations. The vast majority of AI systems currently used or likely to be used in the EU fall into this category. Voluntarily, providers of those systems may choose to apply the requirements for trustworthy AI and adhere to voluntary codes of conduct.

High-risk: A limited number of AI systems defined in the proposal, potentially creating an adverse impact on people's safety or their fundamental rights (as protected by the EU Charter of Fundamental Rights), are considered to be high-risk. Annexed to the Act is the list of high-risk AI systems, which can be reviewed to align with the evolution of AI use cases.

These also include safety components of products covered by sectorial Union legislation. They will always be considered high-risk when subject to third-party conformity assessment under that sectorial legislation.

Unacceptable risk: A very limited set of particularly harmful uses of AI that contravene EU values because they violate fundamental rights and will therefore be banned:

Social scoring for public and private purposes;

Exploitation of vulnerabilities of persons, use of subliminal techniques;

Real-time remote biometric identification in publicly accessible spaces by law enforcement, subject to narrow exceptions (see below);

Biometric categorisation of natural persons based on biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs or sexual orientation. Filtering of datasets based on biometric data in the area of law enforcement will still be possible;

Individual predictive policing;

Emotion recognition in the workplace and education institutions, unless for medical or safety reasons (i.e. monitoring the tiredness levels of a pilot);

Untargeted scraping of internet or CCTV for facial images to build-up or expand databases.

Specific Transparency risk: For certain AI systems specific transparency requirements are imposed, for example where there is a clear risk of manipulation (e.g. via the use of chatbots). Users should be aware that they are interacting with a machine.

In addition, the AI Act considers systemic risks which could arise from general-purpose AI models, including large generative AI models. These can be used for a variety of tasks and are becoming the basis for many AI systems in the EU. Some of these models could carry systemic risks if they are very capable or widely used. For example, powerful models could cause serious accidents or be

misused for far-reaching cyberattacks. Many individuals could be affected if a model propagates harmful biases across many applications.

How do I know whether an AI system is high-risk?

Together with a clear definition of high-risk', the Act sets out a solid methodology that helps identifying high-risk AI systems within the legal framework. This aims to provide legal certainty for businesses and other operators.

The risk classification is based on the intended purpose of the AI system, in line with the existing EU product safety legislation. It means that the classification of the risk depends on the function performed by the AI system and on the specific purpose and modalities for which the system is used.

Annexed to the Act is a list of use cases which are considered to be high-risk. The Commission will ensure that this list is kept up to date and relevant. Systems on the high-risk list, that perform narrow procedural tasks, improve the result of previous human activities, do not influence human decisions or do purely preparatory tasks are not considered high-risk. However, an AI system shall always be considered high-risk if it performs profiling of natural persons.

What are the obligations for providers of high-risk AI systems?

Before placing a high-risk AI system on the EU market or otherwise putting it into service, providers must subject it to a conformity assessment. This will allow them to demonstrate that their system complies with the mandatory requirements for trustworthy AI (e.g. data quality, documentation and traceability, transparency, human oversight, accuracy, cybersecurity and robustness). This assessment has to be repeated if the system or its purpose are substantially modified.

AI systems being safety components of products covered by sectorial Union legislation will always be deemed high-risk when subject to third-party conformity assessment under that sectorial legislation. Also, for biometric systems a third-party conformity assessment is always required.

Providers of high-risk AI systems will also have to implement quality and risk management systems to ensure their compliance with the new requirements and minimise risks for users and affected persons, even after a product is placed on the market.

High-risk AI systems that are deployed by public authorities or entities acting on their behalf will have to be registered in a public EU database, unless those systems are used for law enforcement and migration. The latter will have to be registered in a non-public part of the database that will be only accessible to relevant supervisory authorities.

Market surveillance authorities will support post-market monitoring through audits and by offering providers the possibility to report on serious incidents or breaches of fundamental rights obligations of which they have become aware. Any market surveillance authority may authorise placing on the market of specific high-risk AI for exceptional reasons.

In case of a breach, the requirements will allow national authorities to have access to the information needed to investigate whether the use of the AI system complied with the law.

What are examples for high-risk use cases as defined in Annex III?

Certain critical infrastructures for instance in the fields of road traffic and the supply of water, gas, heating and electricity;

Education and vocational training, e.g. to evaluate learning outcomes and steer the learning process and monitoring of cheating;

Employment, workers management and access to self-employment, e.g. to place targeted job advertisements, to analyse and filter job applications, and to evaluate candidates;

Access to essential private and public services and benefits (e.g. healthcare), creditworthiness evaluation of natural persons, and risk assessment and pricing in relation to life and health insurance;

Certain systems used in the fields of law enforcement, border control, administration of justice and democratic processes;

Evaluation and classification of emergency calls;

Biometric identification, categorisation and emotion recognition systems (outside the prohibited categories);

Recommender systems of very large online platforms are not included, as they are already covered in other legislation (DMA/DSA).

How are general-purpose AI models being regulated?

General-purpose AI models, including large generative AI models, can be used for a variety of tasks.

Individual models may be integrated into a large number of AI systems.

It is important that a provider wishing to build upon a general-purpose AI model has all the necessary information to make sure its system is safe and compliant with the AI Act.

Therefore, the AI Act obliges providers of such models to disclose certain information to downstream system providers. Such transparency enables a better understanding of these models.

Model providers additionally need to have policies in place to ensure that that they respect copyright law when training their models.

In addition, some of these models could pose systemic risks, because they are very capable or widely used.

For now, general purpose AI models that were trained using a total computing power of more than 10^25 FLOPs are considered to carry systemic risks, given that models trained with larger compute tend to be more powerful. The AI Office (established within the Commission) may update this threshold in light of technological advances, and may furthermore in specific cases designate other models as such based on further criteria (e.g. number of users, or the degree of autonomy of the model).

Providers of models with systemic risks are therefore mandated to assess and mitigate risks, report serious incidents, conduct state-of-the-art tests and model evaluations, ensure cybersecurity and provide information on the energy consumption of their models.

For this, they are asked to engage with the European AI Office to draw up Codes of Conduct as the central tool to detail out the rules in cooperation with other experts. A scientific panel will play a central role in overseeing general-purpose AI models.

Why is 10^25 FLOPs an appropriate threshold for GPAI with systemic risks?
This threshold captures the currently most advanced GPAI models, namely OpenAI's GPT-4 and likely Google DeepMind's Gemini.

The capabilities of the models above this threshold are not yet well enough understood. They could pose systemic risks, and therefore it is reasonable to subject their providers to the additional set of obligations.

FLOP is a first proxy for model capabilities, and the exact FLOP threshold can be updated upwards or downwards by the European AI Office, e.g. in the light of progress in objectively measuring model capabilities and of developments in the computing power needed for a given performance level.

The AI Act can be amended to update the FLOP threshold (by means of a delegated act).

Is the AI Act future-proof?
The Regulation introduces different level of risks and provides clear definitions, including for GPAI.

The legislation sets result-oriented requirements for high-risk AI systems but leaves the concrete technical solutions and operationalisation primarily to industry-driven standards that will ensure that the legal framework is flexible to be adapted to different use cases and to enable new technological solutions.

In addition, the AI Act can be amended by delegated and implementing acts, including to update the FLOP threshold (delegated act), to add criteria for classifying the GPAI models as presenting systemic risks (delegated act), to amend modalities to establish regulatory sandboxes and elements of the real-world testing plan (implementing acts).

How does the AI Act regulate biometric identification?
The use of real-time remote biometric identification in publicly accessible spaces (i.e. facial recognition using CCTV) for law enforcement purposes is prohibited, unless used in one of the following cases:

Law enforcement activities related to 16 specified crimes;
Targeted search for specific victims, abduction, trafficking and sexual exploitation of human beings, and missing persons; or
The prevention of threat to the life or physical safety of persons or response to the present or foreseeable threat of a terror attack.

The list of the 16 crimes contains:

Terrorism;
Trafficking in human beings;
Sexual exploitation of children and child sexual abuse material;
Illicit trafficking in narcotic drugs and psychotropic substances;
Illicit trafficking in weapons, munitions and explosives;
Murder;
Grievous bodily injury;
Illicit trade in human organs and tissue;
Illicit trafficking in nuclear or radioactive materials;
Kidnapping, illegal restraint and hostage-taking;
Crimes within the jurisdiction of the International Criminal Court;
Unlawful seizure of aircraft/ships;
Rape;
Environmental crime;
Organised or armed robbery;
Sabotage, participation in a criminal organisation involved in one or more crimes listed above.
Real-time remote biometric identification by law enforcement authorities would be subject to prior authorisation by a judicial or independent administrative authority whose decision is binding. In case of urgency, authorisation can be done within 24 hours; if the authorisation is rejected all data and output needs to be deleted.

It would need to be preceded by prior fundamental rights impact assessment and should be notified to the relevant market surveillance authority and the data protection authority. In case of urgency, the use of the system may be commenced without the registration.

Usage of AI systems for post remote biometric identification (identification of persons in previously collected video material) of persons under investigation requires prior authorisation by a judicial authority or an independent administrative authority, and notification of the data protection and market surveillance authority.

Why are particular rules needed for remote biometric identification?

Biometric identification can take different forms. It can be used for user authentication i.e. to unlock a smartphone or for verification/authentication at border crossings to check a person's identity against his/her travel documents (one-to-one matching).

Biometric identification could also be used remotely, for identifying people in a crowd, where for example an image of a person is checked against a database (one-to-many matching).

Accuracy of systems for facial recognition can vary significantly based on a wide range of factors, such as camera quality, light, distance, database, algorithm, and the subject's ethnicity, age or gender. The same applies for gait and voice recognition and other biometric systems. Highly advanced systems are continuously reducing their false acceptance rates.

While a 99% accuracy rate may sound good in general, it is considerably risky when the result leads to the suspicion of an innocent person. Even a 0.1% error rate is a lot if it concerns tens of thousands of people.

How do the rules protect fundamental rights?
There is already a strong protection for fundamental rights and for non-discrimination in place at EU and Member State level, but complexity and opacity of certain AI applications (black boxes') pose a problem.

A human-centric approach to AI means to ensure AI applications comply with fundamental rights legislation. Accountability and transparency requirements for the use of high-risk AI systems, combined with improved enforcement capacities, will ensure that legal compliance is factored in at the development stage.

Where breaches occur, such requirements will allow national authorities to have access to the information needed to investigate whether the use of AI complied with EU law.

Moreover, the AI Act requires that deployers that are bodies governed by public law or private operators providing public services and operators providing high-risk systems to conduct a fundamental rights impact assessment.

What is a fundamental rights impact assessment? Who has to conduct such an assessment, and when?

The use of a high-risk AI system may produce an impact on fundamental rights. Therefore, deployers that are bodies governed by public law or private operators providing public services, and operators providing high-risk systems shall perform an assessment of the impact on fundamental rights and notify the national authority of the results.

The assessment shall consist of a description of the deployer's processes in which the high-risk AI system will be used, of the period of time and frequency in which the high-risk AI system is intended to be used, of the categories of natural persons and groups likely to be affected by its use in the specific context, of the specific risks of harm likely to impact the affected categories of persons or group of persons, a description of the implementation of human oversight measures and of measures to be taken in case of the materialization of the risks.

If the provider already met this obligation through the data protection impact assessment, the fundamental rights impact assessment shall be conducted in conjunction with that data protection impact assessment.

How does this regulation address racial and gender bias in AI?

It is very important that AI systems do not create or reproduce bias. Rather, when properly designed and used, AI systems can contribute to reduce bias and existing structural discrimination, and thus lead to more equitable and non-discriminatory decisions (e.g. in recruitment).

The new mandatory requirements for all high-risk AI systems will serve this purpose. AI systems must be technically robust to guarantee that the technology is fit for purpose and false positive/negative results are not disproportionately affecting protected groups (e.g. racial or ethnic origin, sex, age etc.).

High-risk systems will also need to be trained and tested with sufficiently representative datasets to minimise the risk of unfair biases embedded in the model and ensure that these can be addressed through appropriate bias detection, correction and other mitigating measures.

They must also be traceable and auditable, ensuring that appropriate documentation is kept,

including of the data used to train the algorithm that would be key in ex post investigations.

Compliance system before and after they are placed on the market will have to ensure these systems are regularly monitored and potential risks are promptly addressed.

When will the AI Act be fully applicable?

Following its adoption by the European Parliament and the Council, the AI Act shall enter into force on the twentieth day following that of its publication in the official Journal. It will be fully applicable 24 months after entry into force, with a graduated approach as follows:

6 months after entry into force, Member States shall phase out prohibited systems;

12 months: obligations for general purpose AI governance become applicable;

24 months: all rules of the AI Act become applicable including obligations for high-risk systems defined in Annex III (list of high-risk use cases);

36 months: obligations for high-risk systems defined in Annex II (list of Union harmonisation legislation) apply.

How will the AI Act be enforced?

Member States hold a key role in the application and enforcement of this Regulation. In this respect, each Member State should designate one or more national competent authorities to supervise the application and implementation, as well as carry out market surveillance activities.

To increase efficiency and to set an official point of contact with the public and other counterparts, each Member State should designate one national supervisory authority, which will also represent the country in the European Artificial Intelligence Board.

Additional technical expertise will be provided by an advisory forum, representing a balanced selection of stakeholders, including industry, start-ups, SMEs, civil society and academia.

In addition, the Commission will establish a new European AI Office, within the Commission, which will supervise general-purpose AI models, cooperate with the European Artificial Intelligence Board and be supported by a scientific panel of independent experts.

Why is a European Artificial Intelligence Board needed and what will it do?

The European Artificial Intelligence Board comprises high-level representatives of competent national supervisory authorities, the European Data Protection Supervisor, and the Commission. Its role is to facilitate a smooth, effective and harmonised implementation of the new AI Regulation.

The Board will issue recommendations and opinions to the Commission regarding high-risk AI systems and on other aspects relevant for the effective and uniform implementation of the new rules. Finally, it will also support standardisation activities in the area.

What are the tasks of the European AI Office?

The AI Office has as its mission to develop Union expertise and capabilities in the field of artificial intelligence and to contribute to the implementation of Union legislation of artificial intelligence in a centralised structure.

In particular, the AI Office shall enforce and supervise the new rules for general purpose AI models. This includes drawing up codes of practice to detail out rules, its role in classifying models with systemic risks and monitoring the effective implementation and compliance with the Regulation. The latter is facilitated by the powers to request documentation, conduct model evaluations, investigate upon alerts and request providers to take corrective action.

The AI Office shall ensure coordination regarding artificial intelligence policy and collaboration between involved Union institutions, bodies and agencies as well as with experts and stakeholders. In particular, it will provide a strong link with the scientific community to support the enforcement, serve as international reference point for independent experts and expert organisations and facilitate exchange and collaboration with similar institutions across the globe.

What is the difference between the AI Board, AI Office, Advisory Forum and Scientific Panel of independent experts?

The AI Board has extended tasks in advising and assisting the Commission and the Member States.

The AI Office is to be established within the Commission and shall work to develop Union expertise and capabilities in the field of artificial intelligence and to contribute to the implementation of Union legislation of artificial intelligence. Particularly, the AI Office shall enforce and supervise the new rules for general purpose AI models.

The Advisory Forum will consist of a balanced selection of stakeholders, including industry, start-ups, SMEs, civil society and academia. It shall be established to advise and provide technical expertise to the Board and the Commission, with members appointed by the Board among stakeholders.

The Scientific Panel of independent experts supports the implementation and enforcement of the Regulation as regards GPAI models and systems, and the Member States would have access to the pool of experts.

What are the penalties for infringement?

When AI systems are put on the market or in use that do not respect the requirements of the Regulation, Member States will have to lay down effective, proportionate and dissuasive penalties, including administrative fines, in relation to infringements and communicate them to the Commission.

The Regulation sets out thresholds that need to be taken into account:

Up to 35m or 7% of the total worldwide annual turnover of the preceding financial year (whichever is higher) for infringements on prohibited practices or non-compliance related to requirements on data;

Up to 15m or 3% of the total worldwide annual turnover of the preceding financial year for non-compliance with any of the other requirements or obligations of the Regulation, including infringement of the rules on general-purpose AI models;

Up to 7.5m or 1.5% of the total worldwide annual turnover of the preceding financial year for the supply of incorrect, incomplete or misleading information to notified bodies and national competent authorities in reply to a request;

For each category of infringement, the threshold would be the lower of the two amounts for SMEs and the higher for other companies.

In order to harmonise national rules and practices in setting administrative fines, the Commission, counting on the advice of the Board, will draw up guidelines.

As EU Institutions, agencies or bodies should lead by example, they will also be subject to the rules and to possible penalties; the European Data Protection Supervisor will have the power to impose

fines to them.

What can individuals do that are affected by a rule violation?
The AI Act foresees a right to lodge a complaint with a national authority. On this basis national authorities can launch market surveillance activities, following the procedures of the market surveillance regulations.

Additionally, the proposed AI Liability Directive aims to provide persons seeking compensation for damage caused by high-risk AI systems with effective means to identify potentially liable persons and obtain relevant evidence for a damage claim. For this purpose, the proposed Directive provides for the disclosure of evidence about specific high-risk AI systems that are suspected of having caused damage.

Moreover, the revised Product Liability Directive will ensure that compensation is available to individuals who suffer death, personal injury or property damage that is caused by a defective product in the Union and clarify that AI systems and products that integrate AI systems are also covered by existing rules.

How do the voluntary codes of conduct for high-risk AI systems work?
Providers of non-high-risk applications can ensure that their AI system is trustworthy by developing their own voluntary codes of conduct or adhering to codes of conduct adopted by other representative associations.

These will apply simultaneously with the transparency obligations for certain AI systems.

The Commission will encourage industry associations and other representative organisations to adopt voluntary codes of conduct.

How do the codes of practice for general purpose AI models work?
The Commission invites providers of general-purpose AI models and other experts to jointly work on a code of practice.

Once developed and approved for this purpose, these codes can be used by the providers of

general-purpose AI models to demonstrate compliance with the relevant obligations from the AI Act, following the example of the GDPR.

This is especially relevant to detail out the rules for providers of general-purpose AI model with systemic risks, to ensure future-proof and effective rules for risk assessment and mitigation as well as other obligations.

Does the AI Act contain provisions regarding environmental protection and sustainability?
The objective of the AI proposal is to address risks to safety and fundamental rights, including the fundamental right to a high-level environmental protection. Environment is also one of the explicitly mentioned and protected legal interests.

The Commission is asked to request European standardisation organisations a standardisation deliverable on reporting and documentation processes to improve AI systems resource performance, such as reduction of energy and other resources consumption of the high-risk AI system during its lifecycle, and on energy efficient development of general-purpose AI models.

Furthermore, the Commission by two years after the date of application of the Regulation and every four years thereafter, is asked to submit a report on the review of the progress on the development of standardisation deliverables on energy efficient development of general-purpose models and asses the need for further measures or actions, including binding measures or actions.

In addition, providers of general purpose AI models, which are trained on large data amounts and therefore prone to high energy consumption, are required to disclose energy consumption.

The Commission is asked to develop an appropriate methodology for this assessment.

In case of general purpose AI models with systemic risks, energy efficiency furthermore needs to be assessed.

How can the new rules support innovation?
The regulatory framework can enhance the uptake of AI in two ways. On the one hand, increasing users' trust will increase the demand for AI used by companies and public authorities. On the other

hand, by increasing legal certainty and harmonising rules, AI providers will access bigger markets, with products that users and consumers appreciate and purchase. Rules will apply only where strictly needed and in a way that minimises the burden for economic operators, with a light governance structure.

The AI Act further enables the creation of regulatory sandboxes and real world testing, which provide a controlled environment to test innovative technologies for a limited time, thereby fostering innovation by companies, SMEs and start-ups in compliance with the AI Act. These, together with other measures such as the additional Networks of AI Excellence Centres and the Public-Private Partnership on Artificial Intelligence, Data and Robotics, and access to Digital Innovation Hubs and Testing and Experimentation Facilities will help build the right framework conditions for companies to develop and deploy AI.

Real world testing of High-Risk AI systems can be conducted for a maximum of 6 months (which can be prolonged by another 6 months). Prior to testing, a plan needs to be drawn up and submitted it to the market surveillance authority, which has to approve of the plan and specific testing conditions, with default tacit approval if no answer has been given within 30 days. Testing may be subject to unannounced inspections by the authority.

Real world testing can only be conducted given specific safeguards, e.g. users of the systems under real world testing have to provide informed consent, the testing must not have any negative effect on them, outcomes need to be reversible or disregardable, and their data needs to be deleted after conclusion of the testing. Special protection is to be granted to vulnerable groups, i.e. due to their age, physical or mental disability.

Besides the AI Act, how will the EU facilitate and support innovation in AI?

The EU's approach to Artificial Intelligence is based on excellence and trust, aiming to boost research and industrial capacity while ensuring safety and the protection of fundamental rights. People and businesses should be able to enjoy the benefits of AI while feeling safe and protected. The European AI Strategy aims at making the EU a world-class hub for AI and ensuring that AI is human-centric and trustworthy. In April 2021, the Commission presented its AI package, including: (1) a review of the Coordinated Plan on Artificial Intelligence and (2) its proposal for a regulation laying down harmonised rules on AI.

With the Coordinated Plan on AI the European Commission has adopted a comprehensive strategy to promote the development and adoption of AI in Europe. It focuses on creating enabling conditions for AI development and uptake, ensuring excellence thrives from the lab to the market, increasing the trustworthiness of AI, and building strategic leadership in high-impact sectors.

The Commission aims to leverage the activities of Member States by coordinating and harmonizing their efforts, to foster a cohesive and synergistic approach towards AI development and adoption. The Commission also put in place the European AI Alliance platform, which brings together stakeholders from academia, industry, and civil society to exchange knowledge and insights on AI policies.

Moreover, the Coordinated plans foresees several measures that aim to unlock data resources, foster critical computing capacity, increase research capacities, support a European network of Testing and Experimentation Facilities (TEFS) and support SMEs through European Digital Innovation Hubs (EDIHs).

What is the international dimension of the EU's approach?
The AI Act and the Coordinated Plan on AI are part of the efforts of the European Union to be a global leader in the promotion of trustworthy AI at international level. AI has become an area of strategic importance at the crossroads of geopolitics, commercial stakes and security concerns.

Countries around the world are choosing to use AI as a way to signal their desires for technical advancement due to its utility and potential. AI regulation is only emerging and the EU will take actions to foster the setting of global AI standards in close collaboration with international partners in line with the rules-based multilateral system and the values it upholds. The EU intends to deepen partnerships, coalitions and alliances with EU partners (e.g. Japan, the US, India, Canada, South Korea, Singapore, or the Latin American and Caribbean region) as well as multilateral (e.g. OECD, G7 and G20) and regional organisations (e.g. Council of Europe).

*Updated on 14/12/2023

Document 7 (Source: https://bg3.wiki/wiki/The_Emperor)

The Emperor is a mind flayer who appears in Baldur's Gate 3. It[note 1] plays a key role in the main

story, but its identity is intentionally obscured until later parts of the game, allowing the player to ultimately decide for themselves if they want to know more about it, and whether or not it is trustworthy.

Contents

Overview

Identity

The Emperor plays a key role in the main story of Baldur's Gate 3, and as part of this role its identity and personal background are kept obfuscated for much of the game. It very carefully divulges

information that it deems necessary, sometimes arguing that the player is not ready for the answer yet, or that it will reveal specific information in the future.

During Acts One and Two, the Emperor only "meets" with the player as the Dream Guardian. At the beginning of Act Three, the player finally meets the Emperor face to face, an event which reveals that it is a mind flayer.

Through all three Acts, the Emperor generally serves as a guide, and unlikely ally to the party, having the means to protect their minds from the influence of the Absolute, through the use of the prisoner within the Astral Prism.

"Don't let my form deceive you. I am the one that's been protecting you. I am the one that came to you in your dreams. Help me.
The Emperor, during Act 3

Personal quest

After reaching the Elfsong Tavern in Act Three, the Emperor will initiate the quest Visit the Emperor's Old Hideout, in which the player can better get to know the Emperor. It discloses some of its past, during its time in the city and from before it became illithid.

Recruitment

The Emperor can appear in multiple combat encounters as a controllable ally, a neutral ally, or an enemy. It cannot, however, become a full member of the player's party or camp.

Romance

The Emperor can have a romance with the player during Act Three. See Romance.

History

Details about the Emperor's personal history are intentionally obfuscated during most of the game, but the player has the opportunity to learn more about it through conversations, interactions with other characters, reading books, and completing specific side quests.

Ico knownSpells lvl 03.png Act 3 Spoilers! This section reveals details about the story of Baldur's Gate 3.

An Adventurer, I came from Baldur's Gate, though I was never one to be constrained by circumstance. I longed for more.

That longing brought me to Moonrise Towers on a search for treasure. To a colony of mind flayers who caught me and changed me.
The Emperor was once Balduran, an adventurer who founded a coastal village called Grey Harbour. After securing enough money to fund the building of the Wall that led to Baldur's Gate being founded, he felt the call of the sea once more. On the voyage, and following a shipwreck, Balduran made his way to Moonrise Towers in search of fortune. There, he found a coven of mind flayers who infected him with an illithid tadpole. As a record of his interrogation by Enver Gortash during the planning phases of the Absolute Hoax states, he spent ten years under the thrall of the Moonrise Elder Brain.

After Balduran was reborn as an illithid and broke free from the Elder Brain the Absolute, it returned to Baldur's Gate, living in the shadows and feeding on the brains of criminals. Initially struggling with its identity as a mind flayer, Balduran eventually embraced its new form.

Balduran's new acceptance of its illithid form caused a wedge to form between it and its close companion, the dragon Ansur. Ansur attempted to kill Balduran as it slept, believing this would be a merciful death. The Emperor sensed the attempt, and in its struggle to protect itself from being murdered, it killed Ansur in self-defence. [1]

After Ansur's death, Balduran came to be called the Emperor as it used its newfound psychic influence to rule Baldur's Gate from the shadows. For the next four centuries, it made its haven under the Elfsong tavern, keeping various sentimental knick knacks from its time as Balduran.

I had the fortune of meeting Duke Stelmane. We formed a partnership
During those four centuries, it also came to be associated with the Knights of the Shield, a lawful and neutral evil conglomerate of politicians and merchants manipulating events behind the scenes. Duke Stelmane was a major figure of this secret society, acting as the Emperor's envoy while it

secretly kept her enthralled. [note 2]

Sometime before the events of the game, Enver Gortash and the Dark Urge captured the Emperor, and brought it back under the thrall of the Moonrise Elder Brain, who was now wearing the Crown of Karsus and had become the Netherbrain masquerading as the Absolute. The Netherbrain, sought to have all three Chosen of the Dead Three killed, and specifically picked the Emperor, unbeknown to it, to lead a team of illithids on a nautiloid to search for and steal from the Githyanki the Astral Prism containing their prince, Orpheus.[2]

Events of Baldur's Gate 3

Act Two finale

Main article: Help Your Protector

On the way to Baldur's Gate, the party will be ambushed by a group of Gish'ra warriors while resting at Wyrm's Lookout. Entering the portal to the Astral Prism, the party will hear their Dream Guardian calling out for help. However, when the party reaches them, it is only to discover that the true identity of their visitor is the illithid known as the Emperor.

After defending the Emperor, it will explain how it used the power of the Prism and Orpheus to protect the party from the Absolute, and recite to the party its history as an adventurer and finding freedom from the Absolute. The Emperor will offer the party an Astral Touched Tadpole, which causes the user to transform into a partial-illithid. It insists the path of the mind flayer is preferable, regardless of the player's view on them.

Though this may seen contradictory to its previous promise as the Dream Guardian; to ensure the party do not become mind flayers, this promise refers to the player becoming a mind flayer unwillingly because of the Elder Brain. The Emperor is in favour of the player becoming a mind flayer of their own volition and without the influence of the Elder Brain.

Act Three

Elfsong Tavern

Main article: Visit the Emperor's Old Hideout

As the party nears the Elfsong, the Emperor will remark that the tavern is the location of its old hideout. The hideout proper is in the basement, past the Knights of the Shield's hideout. In it, the player will find various sentimental knick knacks from the Emperor's previous life, before becoming

an illithid.

Around the room is its old dog Rascal's collar, its favourite recipe (fiddlehead soup), its first adventuring sword, and part of a cutlery set from its mother; the butter knife having been lost during its last shipwreck on the Isle of Balduran, inside the wreck of the Wandering Eye ship.

There are also some more illithid-adequate items such as chains for its preferred prey - allegedly criminals and lawbreakers - and jars for brains.

The Wyrmway

See also: Wyrmway and The Blade of Frontiers

Once the party completes the Wyrmway trials, they will find the corpse of Ansur the Dragon. Interacting with his body will awaken Ansur's spirit, which briefly possesses the player in order to communicate. As Ansur's introduction concludes, he will detect the Emperor within the Astral Prism.

Ico knownSpells lvl 03.png Act 3 Spoilers! This section reveals details about the story of Baldur's Gate 3.

Ansur will reveal that the Emperor in fact was formerly Balduran, the founder of Baldur's Gate. Furthermore, he explains that while the Emperor initially did not want to become a mind flayer, it eventually fully embraced its new form, and its comfort with this caused a rift between the Emperor and Ansur. After "exhausting all possibility of reversing (the Emperor's) condition", Ansur was agonizing and the Emperor (as seen in the letter on Ansur's body) attempted to convince him to leave. Ansur then attempted to murder the Emperor during its sleep as a mercy killing, and the Emperor killed Ansur in self-defense.

This development is somewhat foreshadowed when the player first meets The Emperor in their true form, as the song that plays during the encounter is a variation of The Elf Song, which prominently features Balduran in its lyrics.

Endings

Ico knownSpells lvl 03.png Act 3 Spoilers! This section reveals details about the story of Baldur's Gate 3.

Let the Emperor use the Netherstones

The Emperor unless convinced otherwise is mostly concerned with its survival and prosperity. Should the player allow it to wield the Netherstones, it will follow through on destroying the Elder Brain, at the cost of letting it "assimilate" with Orpheus.

If the player suggests to the Emperor to take control of the Netherbrain, it will mention that the thought of becoming the Absolute did cross its mind. But unless otherwise persuaded, it will refuse, claiming that whoever becomes the leader of the Cult of the Absolute will be in an open war with the Githyanki, which is a war it is not certain it will survive. The Emperor will destroy the Netherbrain, and the parasites within its control in this ending.

The Emperor controls the Netherbrain

It is also possible, after suggesting it to take control of the Netherbrain, to persuade it. In this scenario, it does not free the player or their party, instead making them mindless thralls and assuming absolute control of them, continuing the Grand Design.

Orpheus is freed

If the player frees Orpheus, the Emperor will abandon the party, and side with the Netherbrain for the sake of its own survival, as it believes that Orpheus will kill it.

Attack the Emperor

The Emperor can be attacked and killed when it first reveals itself to be a mind flayer. This will result in the influence of the Netherbrain taking over control of the party, ending the game.

List of interactions

See Dream Guardian to read about its previous conversations with the player when it was in disguise.

Charm Person Icon.png Romance Spoilers This section reveals details about romance and may contain mature themes.

Players have a limited number of opportunities to interact with the Emperor, and as such, opportunities for conversation are much more limited compared to that of companions.

Conversation scenes are available, but only occur during Act 3, after its "true" identity is revealed to the player, and all scenes require a long rest to trigger. The Emperor will occasionally also talk to the player as they walk through different locations in Baldur's Gate.

Conversation scenes

Known conversation opportunities with the Emperor currently include the following cases, but each scene appears to have multiple outcomes that affect the tone of all subsequent conversations.

Depending on the player's choices, the Emperor's behaviour has many possible states. The more the player treats the Emperor like a "person", the more it will act as such, compared to other illithids. The more the player treats The Emperor like a monstrosity with hostile intent, the more it will respond to the player with threatening language and visions of it acting like a hostile illithid.

Identity revealed

During Help Your Protector at the start of Act 3, a conversation is automatically triggered when the player ventures far enough into the Astral Plane. A combat encounter in some form is inevitable from this conversation, and then another set of conversation options are available after the combat resolves. The Emperor will have nothing further to say when this conversation ends, even if the player tries to interact with it further.

Regarding Duke Stelmane

When the player first explores the Rivington area, being in proximity to certain characters or objects will "inform" the player about the recent death of Duke Belynne Stelmane. This will trigger a line of ambient commentary from The Emperor. The next time a Long Rest is triggered, the player may trigger a scene discussing The Emperor's reactions in more depth. Certain dialogue choices made during earlier conversations seem to disqualify the player from this scene. If the player does not long rest before completing the quest Visit the Emperor's Old Hideout, this scene will be skipped entirely.

On conclusion of Visit the Emperor's Old Hideout

This scene may be available to trigger (by long resting) after the player completes the quest Visit the Emperor's Old Hideout.

Possible states for this scene appear to vary heavily depending on the player's choices in prior conversation scenes, with the general differentiating factor being the "attitude" the player appears to express towards illithids, and towards the Emperor, through their selected options in these prior scenes.

If the player tried to kill the Emperor in Act One, by choosing the dialogue option "You do a great

impression of a human. But you're not fooling me." , the Emperor offers to share memories through a vision. This vision shows Stelmane paralysed in pain, being brainwashed, and turning into the Emperor's puppet. Her face emotionless, and the Emperor puppeteering her gestures to get a sense of company. Such was its true relationship with Duke Stelmane. [note 2]

The Emperor uses this memory to frighten the player. It gives them orders, and threatens to make them half-illithid even if they refuse.

Romance

In terms of game mechanics, it is technically possible to romance the Emperor. [note 3]

If the player chooses to reject its advances, the Emperor's attitude in conversation will change in a way that appears to be reactively appropriate to the way it was treated. For example, if the option "Absolutely not, you freak!" is chosen at any opportunity, the Emperor's treatment of the player takes a much more hostile tone in all future interactions.

Players have a limited number of opportunities to interact with the Emperor, and as such, opportunities for romantically-styled interactions are much more limited compared to the other primary companions.

If the player visits Crèche Y'llek prior to the start of Act 3, killing the Dream Guardian will subsequently lock the player out of romancing the Emperor, and from interacting with it in general.

There are many possible ways to interact with the Emperor in the available conversation scenes. It currently seems that the primary way to unlock "romantic" options is by choosing dialogue that generally treats the Emperor more like "any other person", and does not show explicit hostility towards its actions, or its illithid characteristics.

The player does not need to accept the powers of the Astral-Touched Tadpole to unlock this option. The Emperor seems to take offence to destroying the tadpole, but more testing is needed to determine if this has any effect on the available scenes.

The scene that occurs after completing Visit the Emperor's Old Hideout is generally regarded as the

"primary" romantic scene. As long as the player is receptive to the Emperor's advances, conclusions to this scene will allow the player to engage in more intimate activities with it.

Conversation options that acknowledge this romance (after the primary scene has concluded) appear to exist in a limited number of places. For example, it is possible to tell Raphael "I don't want any part of this the Emperor is my lover." during a specific conversation, if initiated after the romance scene has happened.

Engaging in the primary scene has no effect on other ongoing romances, even when romancing Lae'zel, who is generally hostile to illithids.

Achievements
A-Mind Blown.jpg
Mind Blown
Romance the Emperor.

Gallery
They called me The Emperor
They called me The Emperor

The Emperor feeding on criminals
The Emperor feeding on criminals

Character portrait by Edward Vanderghote
Character portrait by Edward Vanderghote

The Emperor's model
The Emperor's model

Notes

The Emperor's existence confirms the Dream Guardian as being an illithid influence, albeit in a different way.

In Early Access, the Dream Guardian (known then as Dream Visitor) was implied to be a mental manifestation of the player's tadpole, as it eased them towards using their powers more, as well as showing them a future of domination and control.

In the Full Release, the Emperor plays a similar role, in the sense that it also encourages the player to expand their potential through using the tadpole's power, but it is much more passive. In addition, its interests seem to be aligned against the Absolute.

Footnotes

The Emperor, like other mind flayers, is addressed using the "it" pronoun. It is incidentally referred to as "he" in-game, and "they" in the game's files, possibly due to an oversight, or characters conflating its current and previous identities.

The Emperor's vision of its control over Belynne Stelmane is corroborated by the 5e module, Baldur's Gate: Descent into Avernus. In it, Stelmane is described as having a secret, mental battle against a mind flayer. This mind flayer is very likely the Emperor itself, and as a result, puts its entire "alliance" with Stelmane into question. It is very possible the Emperor and Stelmane did not have a proper alliance at all, and rather, the Emperor enthralled her for its needs. Whether this was always the case, or if they had a genuine alliance beforehand, isn't fully clear.

This romance behaves somewhat differently from that with companions, as the Emperor generally cannot be interacted with outside of cutscenes, and romantic progression is limited to the final act of the game.

References

Dialogue with Ansur.

The Netherbrain's dialogue to the player at the Morphic Pool.