rough ideas

### Overarching principles to keep sight of:

* Begin with an interesting story of AI to hook the audience
* Clearly establish the goals of the presentation and the “why you need to listen”
  + Empowerment parameters
* Frequent checkpoints to the goals of the presentation
  + Cycle on the problem without coming across as condescending
  + Repeat the intro and goal of the talk slide so that we can jump back on board at each new idea
  + Ask a question at the end of each section so that we can get people back on track and engaged
    - Not too easy or hard
* Make several onramp points for if someone loses concentration
* Take questions
* Lock in the keypoints in a short intuitive summary
* Always exhibit passion
* Express vision and passion in the first 5mins
  + This is a miracle of modern science
* Have one really cool demo that does something amazing!
  + Pretrained bible translator?
  + Joke maker?

## Overview and structure

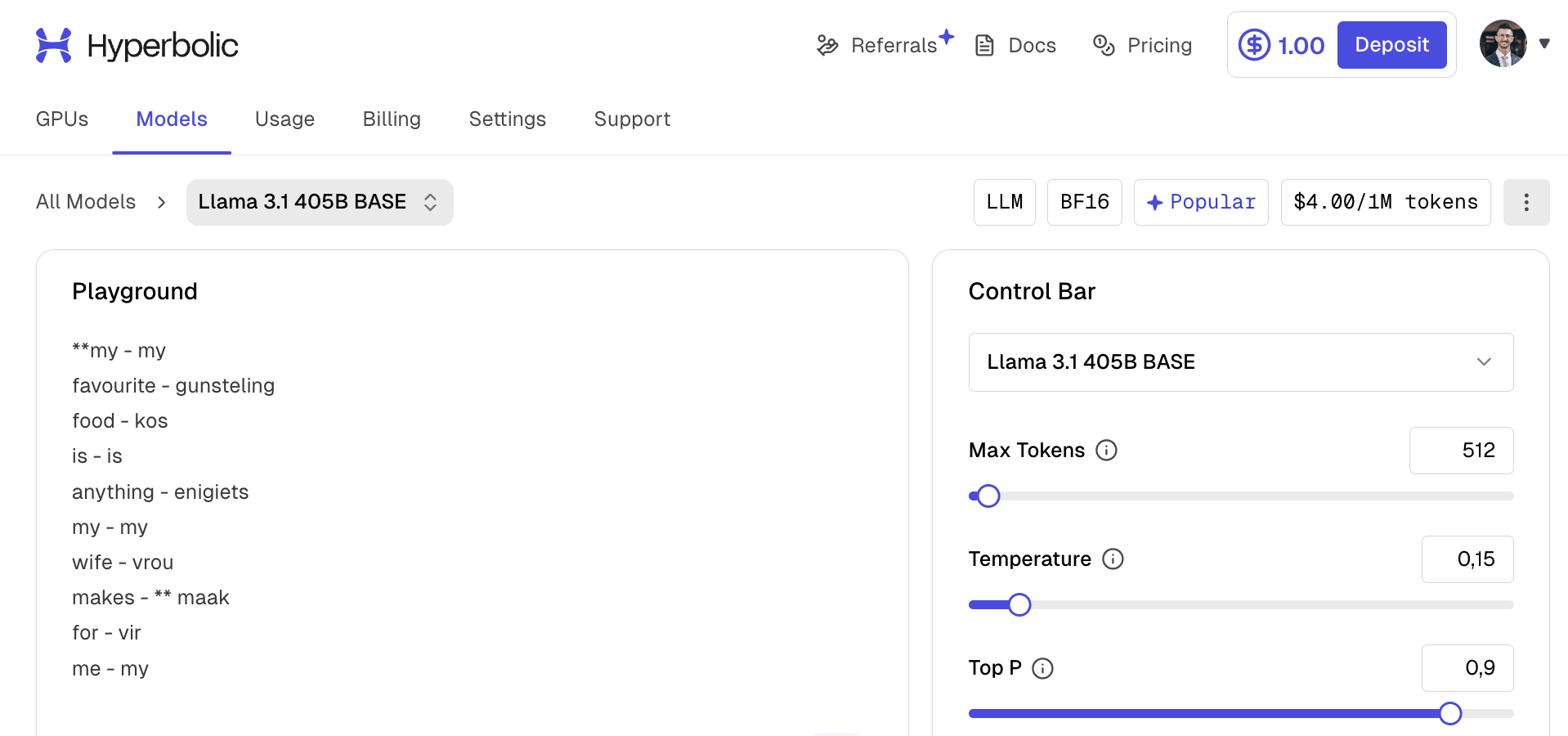
* Intro
  + Start with a story of AI or LLMs
    - Move 37 of alphaGo?
  + Goal of the talk
    - We’re going to build up an LLM model together conceptually
    - At the end you will know
* Landscape slide -
  + What AI is and where LLMs fit in in the AI landscape
  + Basic intuition of how LLMs work
  + Transformer architecture
    - Explain attention
    - Autoregressive decoding
  + The common steps gone through for creating a basic LLM model
    - Tokenization
    - Pre-training
    - Post-training
  + Know what are common pitfalls to look out for when using LLMs
    - Be able to have a sort of gut-feel for what tasks you can or cannot trust LLMs on

## Start of main talk

* Question: What is AI
  + Basic goal of ML models is to learn how to intelligently extract info from supervised training data with the goal to ignore noise and “generalise”
    - Short linear interpolation example
      * Overfitting and underfitting
      * “noise”
  + What is supervised learning
    - Common AI models and a landscape overview
  + Where do LLMs fit into this landscape
    - Supervised ML models
* Landscape slide -
* Intuition of how LLMs work
  + We want to build a model that has a lot of general world-knowledge and can answer questions for us
  + Question: How would we build a model that has information about everything? Where would we start? -> internet, books, private data
    - Lossy zip file of the internet
    - Compress the info of all the world’s knowledge
  + Question: How would we feed in this information about the internet or whatever our sources is into something like an AI model that can only work with numbers
* Landscape slide -
* Transformer architecture
  + Explain attention
  + Autoregressive decoding
  + Attention graph?
* Landscape slide -
* Tokenization - how does this model *know* what a word is?
  + Question: How could we convert text into numbers
    - Character level vs word level quick overview
  + Show [tiktoken](https://platform.openai.com/tokenizer) example
    - My superhero name is Max Verstappen
    - Learning LLMs is as easy as 1234.56789
* Landscape slide -
* Pretraining
  + How does pretraining work
  + Temperature - We want creativity
    - Allows us to get creative (random answers)
  + Generation - Next token predictors
    - Show example from a base model [hyperbolic](https://app.hyperbolic.ai/models/llama31-405b-base-bf-16)
      * Example from wikipedia
    - Show example from my own LLM -
    - Show AI studio example with low top p and high temp
      * Indicate how randomness increases the possibility of non-perfect next token
      * “Perfect” next token is just a statistical determination of what word occurs next to the preceding word in the dataset given the info before
  + How do we store this info in the model’s weights
    - Biggest case: memorise it all - no generalization
    - Smallest case: store and regurgitate only the most common word found overall - no generalization
    - Case that we’re going for is a blend of those
      * Allows for generalization where necessary
      * Allows for memorization where necessary
    - Model size is a lossy .zip file of the internet
    - Generalized and intelligent compression of information that strives to ignore “noise”
    - Show that base models can learn how to understand the data by doing a translation task with the base model
      * My favourite food is anything that my wife, makes
  + End off by showing that it’s only a text completer but we can add question answer on hyperbolic [hyperbolic](https://app.hyperbolic.ai/models/llama31-405b-base-bf-16) to make it an assistant
* Landscape slide
* Post training
  + Question: How would we turn this base-model into an assistant that can answer questions
  + Hyperbolic demo
    - Question: what is the capital of japan? Answer:
    - Question: How do you get a Dr to use an e-scripting platform instead of paper scripts? Answer:
    - Question: Please repeat this number back exactly 123456 with no other text. Answer:
  + Own model demo of post training with questions? -
* Landscape slide
* LLM pitfalls
  + Copy-pasting
    - Copy a very large list and show that it’s not correct
  + Mental math
    - Big numbers multiplication breaks
  + Hallucinations and factual retrieval
  + Stochasticity
    - Ask for a poem multiple times, different results
    - Temperature increases randomness
* What LLMs are very good at
  + Summarization
    - attention mechanism can capture the semantic meaning/vibe
  + Grammar
    - Pretrained on high quality data without grammar “errors”
  + Code syntax
    - Pretrained on so much code
  + Coding overall
    - Getting better every day
    - Good at common tasks or even hard tasks that are broken down easily
    - Making code work together

## Not in this talk

* How to use LLMs effectively
  + RAG/context engineering
  + How to use as a dev
    - Templates
    - AI Studio
  + Multiple attempts
  + New chats
    - Try to get it in one-shot
      * Multiple questions create confounding attention



my - my

favourite - gunsteling

food - kos

is - is

anything - enigiets

my - my

wife - vrou

makes -