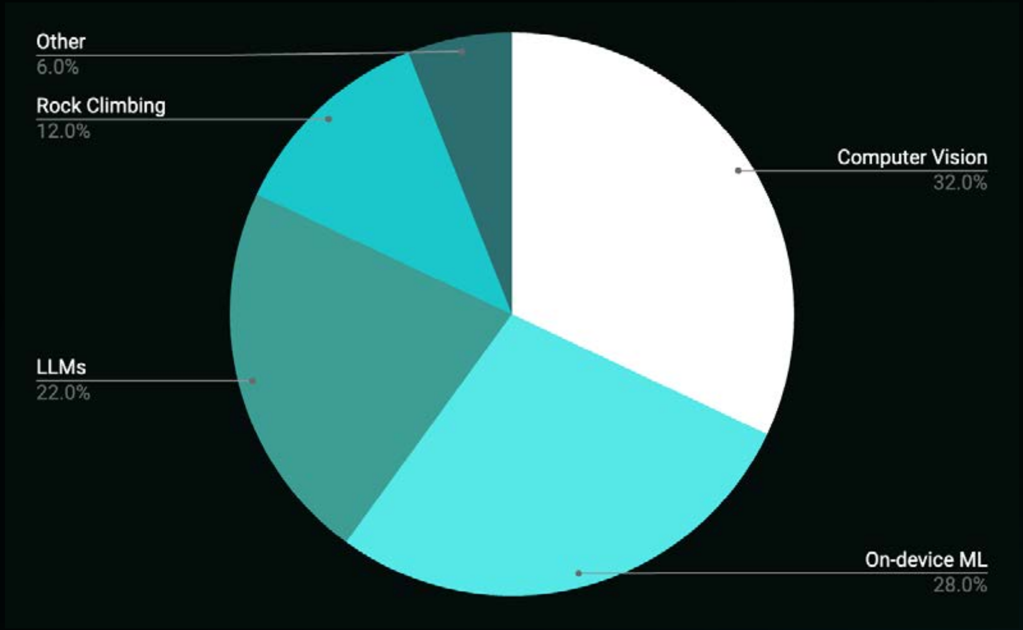
On Device LLMs

A Path to Safer and More Efficient AI Systems

Who am I



- ● Computer Vision
- ● On-device ML
- ● LLMs
- ● Rock Climbing
- ● Other

TABLE OF CONTENTS

01 Why on-device LLMs

02 State of the art

03 Current use cases

04 Path to MVP

05 5 years later

06 Conclusion

01

Why On-Device LLMs matter?

Understanding Large Language Models

THREE WORD GAME

A 2-player game where one person says three words, and the other follows with three words that continue the story. This process is repeated until a story is formed.

LARGE LANGUAGE MODELS

LLMs follow a similar approach. Every time the user provides input, the model generates a continuation, producing clear and relevant responses.

Problems with Large Language Models

PRIVACY INVASION

LLMs can store vast amounts of personal data shared by users during interactions giving companies building LLMs potential access. This access raises concerns about user privacy and the potential for misuse of personal information.

CARBON FOOTPRINT

Recent estimations have shown that training a GPT-4 has carbon emissions equivalent to driving 18 million miles in a gasoline car. Estimations have also shown per query to GPT-4 can be equivalent to charging a mobile phone 60 times.

ABILITY TO MIMIC HUMANS

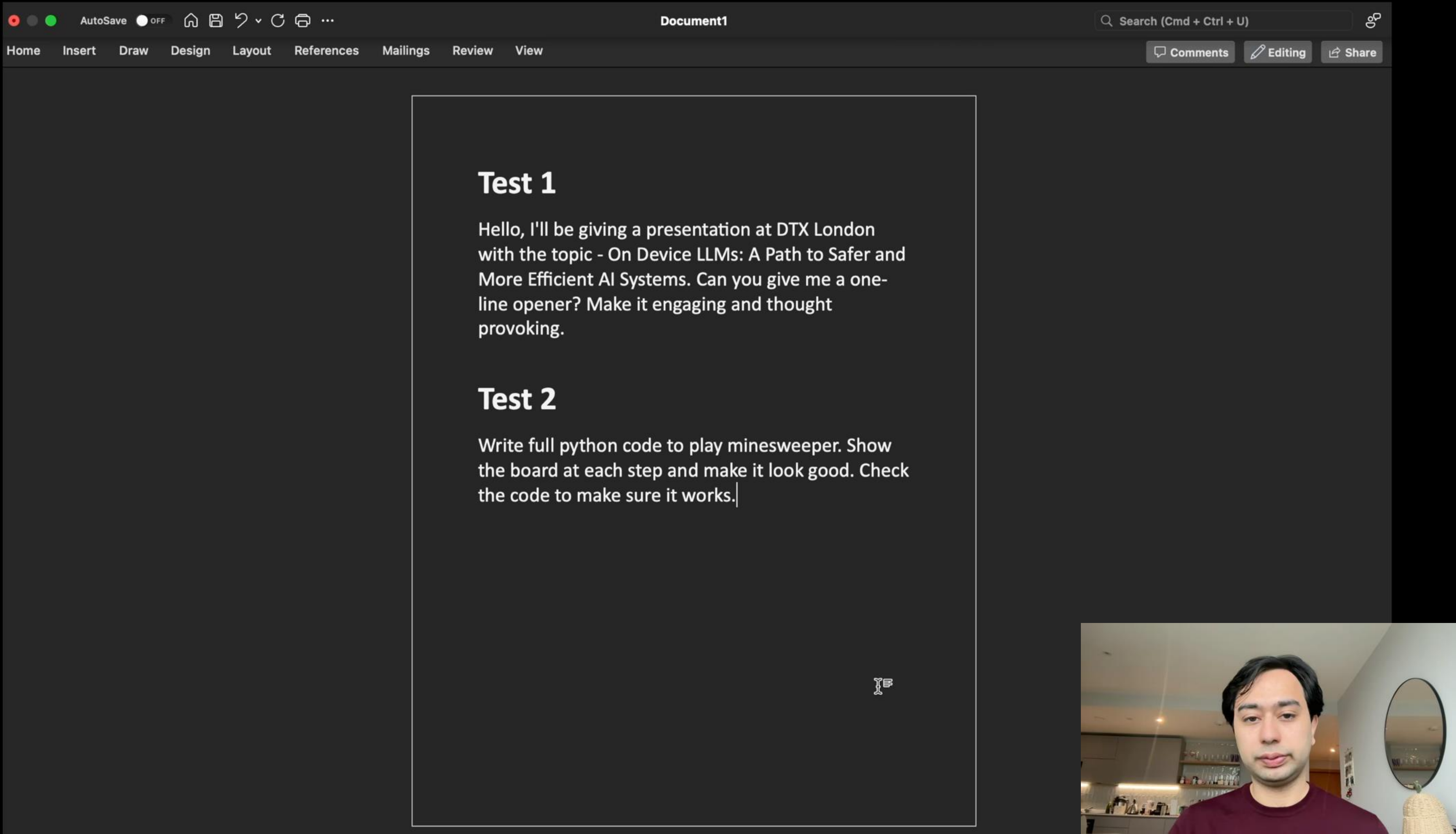
LLMs mimicking humans can be dangerous as they may spread misinformation, create deepfakes, and commit fraud. Their human-like output can also erode trust in online communic
ethical cor

Privacy and Efficiency improvements due to on-device LLMs

1. **On-Device, Off the Grid:** Data stays on your device, with no need to send anything to external servers.
2. **No Middlemen, No Risks:** Eliminates third-party access, significantly reducing breach risks and ensuring your data stays private.
3. **Smaller, Smarter, Greener:** On-device LLMs are designed to be lightweight for limited hardware, using far less energy than the massive data centers needed for cloud models, reducing carbo

02

State of the art On-Device LLMs

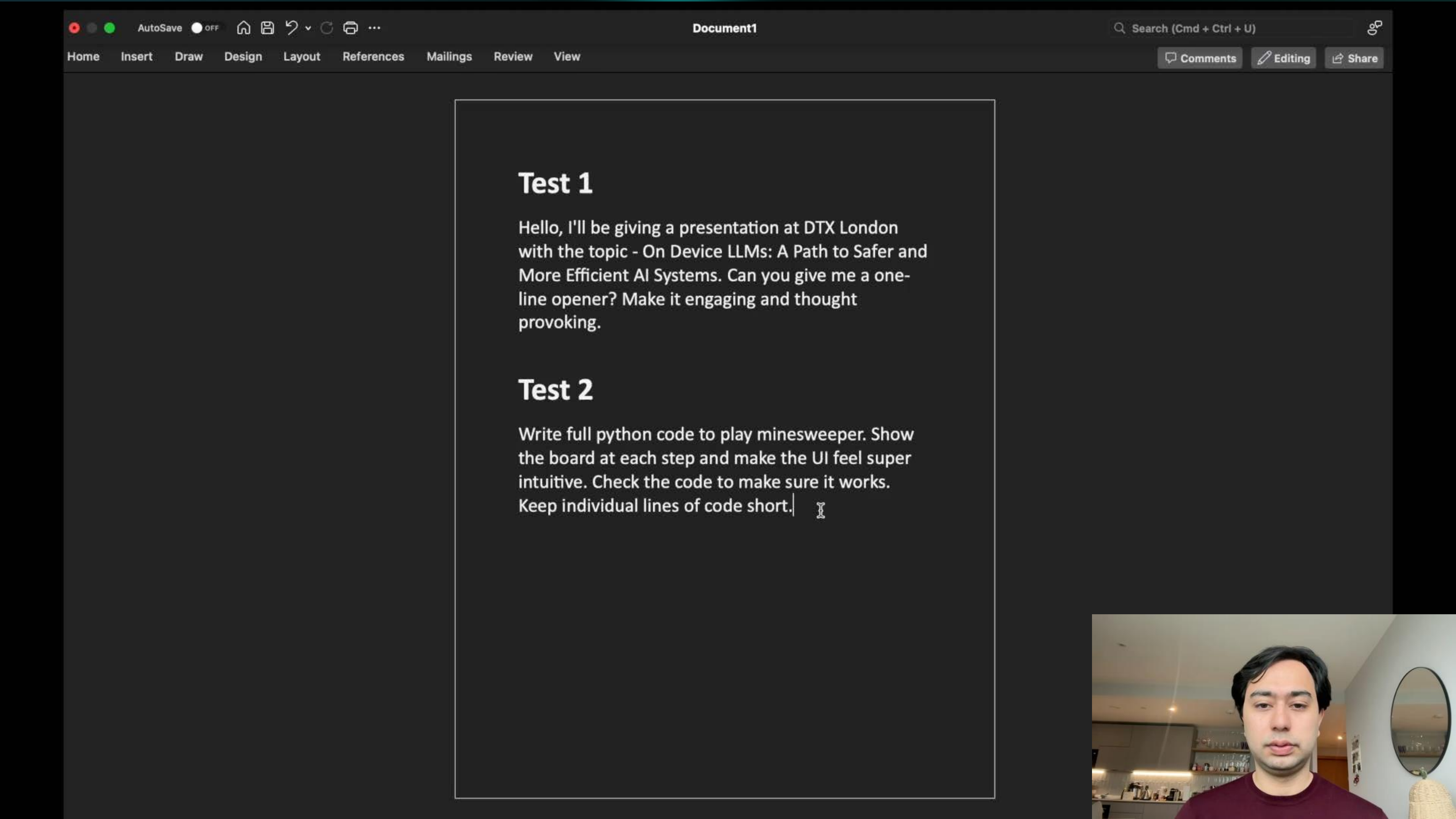


Test 1

Hello, I'll be giving a presentation at DTX London with the topic - On Device LLMs: A Path to Safer and More Efficient AI Systems. Can you give me a one-line opener? Make it engaging and thought provoking.

Test 2

Write full python code to play minesweeper. Show the board at each step and make it look good. Check the code to make sure it works.



Test 1

Hello, I'll be giving a presentation at DTX London with the topic - On Device LLMs: A Path to Safer and More Efficient AI Systems. Can you give me a one-line opener? Make it engaging and thought provoking.

Test 2

Write full python code to play minesweeper. Show the board at each step and make the UI feel super intuitive. Check the code to make sure it works. Keep individual lines of code short.

Benchmarks

Model	Chat / Reasoning (MMLU)	Coding (MBPP EvalPlus)
Llama3.1 (8b)	73.0	72.8
Llama3.1 (70b)	86.0	86.0
GPT 4o	88.7	87.8
CodeQwen1.5 (7b)	N/A	78.7

03

Current Real World Use Cases

Current Use Cases

Style Transfer

On-device LLMs enable seamless style transfer, allowing users to modify the tone or style of their writing. This customization helps adapt text for different contexts, such as shifting from formal to casual.

Speech to text

Although this isn't strictly a function of large language models(LLMs), on-device speech-to-text systems convert spoken language into written text in real-time.

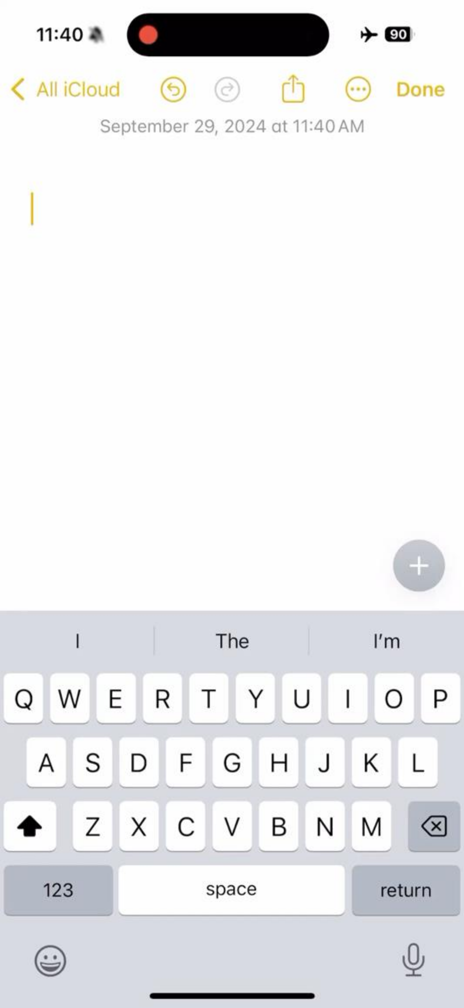
Summarization

On-device LLMs can generate concise, accurate summaries of texts in real-time without internet access. Their efficiency, however, might be limited for very detailed or nuanced content compared to server-based models.

Translation

On-device LLMs can provide accurate, real-time translations without an internet connection, handling multiple languages effectively. However, they may fall short compared to server-based solutions for complex or rare languages.

Apple Intelligence On-Device Features



04

Path to MVP: Reaching GPT4o level perf on device

~30 billion

Parameters need to run on-device

Specialization

Needs to be a layer deeper

Adapters

Need a step-function improvement

Software Innovations

1. **Enhanced Compression Techniques :** Development of advanced algorithms to optimize model size and enable efficient parameter usage for on-device processing.
2. **Application-level Specialization :** Implementation of deeper and more granular specialization tailored to specific applications to improve processing speed and accuracy on-device.
3. **Improved Adapter Mechanisms:** Construction of sophisticated adapter frameworks to facilitate significant functional improvements and seamless integration with existing on-device LLMs.

Hardware Innovations

1. **Optimized Processing Units:**
   Development of specialized chips and processing units designed explicitly for handling large-scale language models efficiently on-device.
2. **Enhanced Memory Architecture:**
   Innovation in memory systems to support rapid access and management of extensive model parameters, ensuring high-speed performance.

05

Impact of on-device LLMs 5 years later

Use cases enabled

Augmented Workflows

Augment productivity with real-time, context-aware support. They streamline tasks like emails, data analysis, and presentations, allowing professionals to focus on creative problem-solving.

On-Device Therapy

Revolutionize therapy by providing real-time emotional support and personalized mental health advice. This ensures privacy and continuous, adaptive care, significantly enhancing accessibility and effectiveness.

Legal Assistants

Transform legal assistance by delivering instant support for drafting documents, analyzing cases, and providing advice. This ensures confidentiality and efficiency, making services more accessible and streamlined.

Medical Diagnosis

Reshape medical diagnostics with instant, intelligent analysis and personalized health advice. This guarantees data privacy and delivers timely, accurate diagnoses, significantly improving healthcare accessibility and quality.

THANKS!

For questions or follow-ups, email me at
rish@pinnacle.co