AMD
RYZEN AI

CONSUMER AI PERFORMANCE
WITH LARGE LANGUAGE MODELS (LLMS)

MARCH 2024

RUNNING AN LLM ON YOUR PROCESSOR
WITH LM STUDIO

LOCAL = PRIVACY | **NO SUBSCRIPTION FEES** | **NO WI-FI NEEDED**

100 Million
users within 2 months
on ChatGPT

Llama 2	Popular LLM
MISTRAL AI_	Comparable to GPT 3.5/4 models
code llama	Used for coding with markdown
RAG	Local context for LLMS

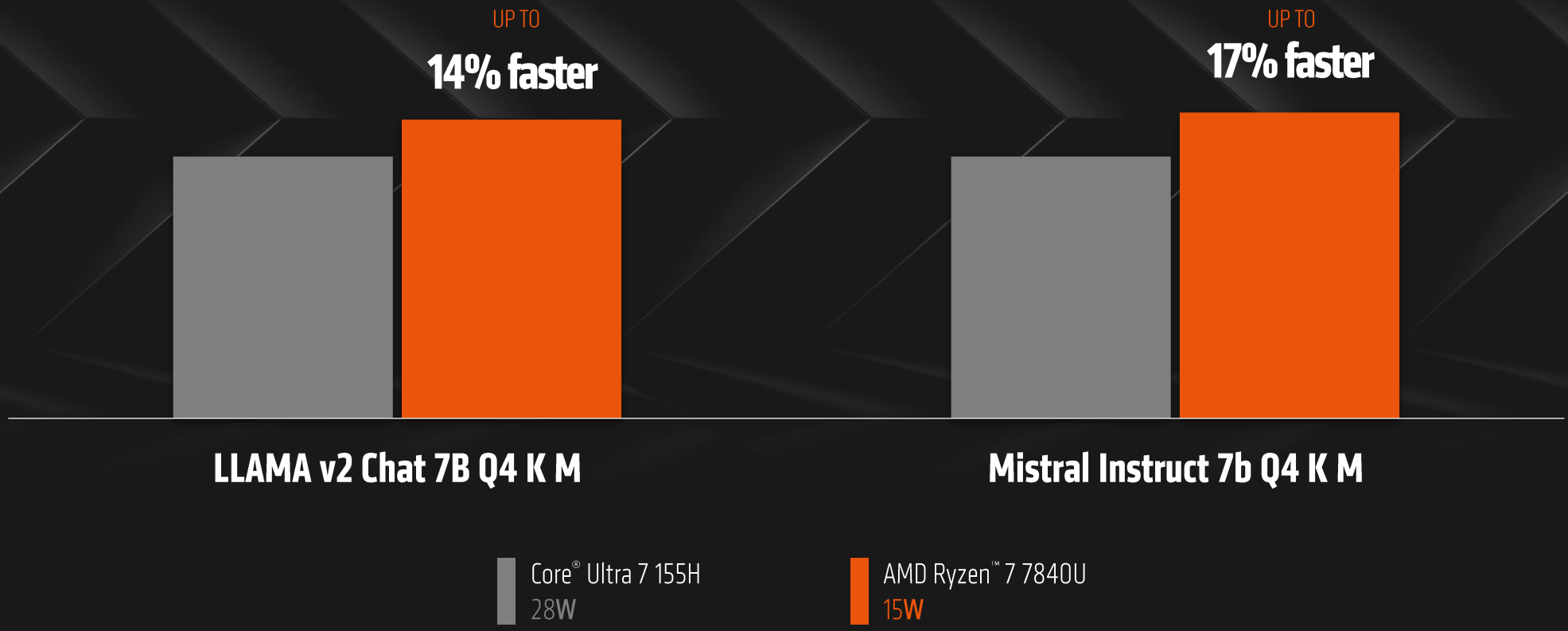
GET YOUR LLM UP *AND RUNNING*

You don’t need to be a software developer to enjoy the benefits of an LLM on your personal PC

AMD
together we advance_AI

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI

TOKENS PER SECOND FOR SAMPLE PROMPT ON LM STUDIO

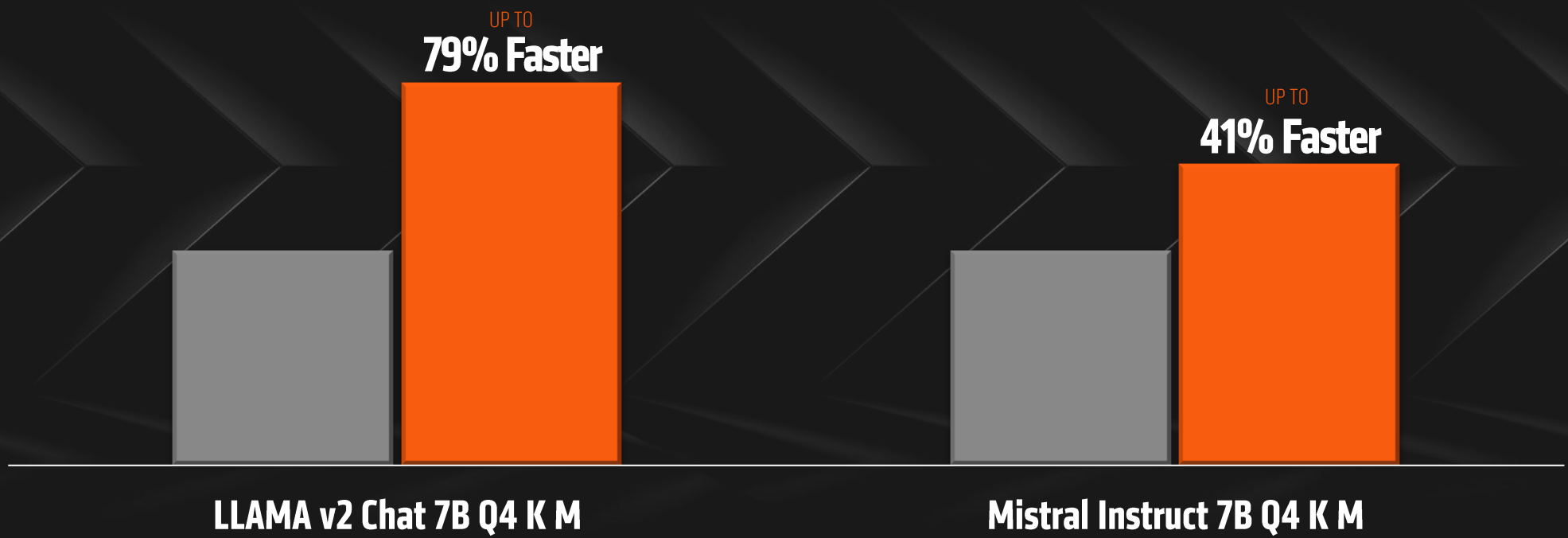


SEE ENDNOTE PHX-59

AMD
together we advance_AI

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI

TIME TO FIRST TOKEN FOR SAMPLE PROMPT ON LM STUDIO



HIGHER IS BETTER

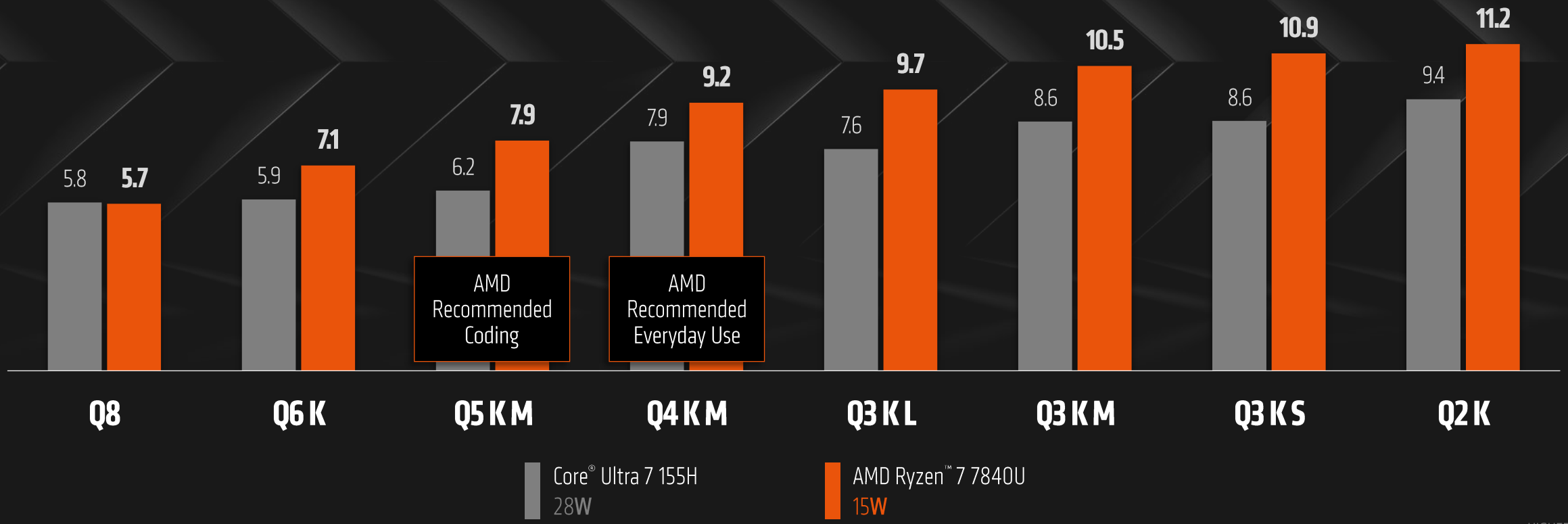
SEE ENDNOTE PHX-59

MISTRAL AI_

MISTRAL INSTRUCT 7B

TOKENS/S WITH QUANTIZATION SCALING ON LM STUDIO

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI



HIGHER IS BETTER

SEE ENDNOTE PHX-59

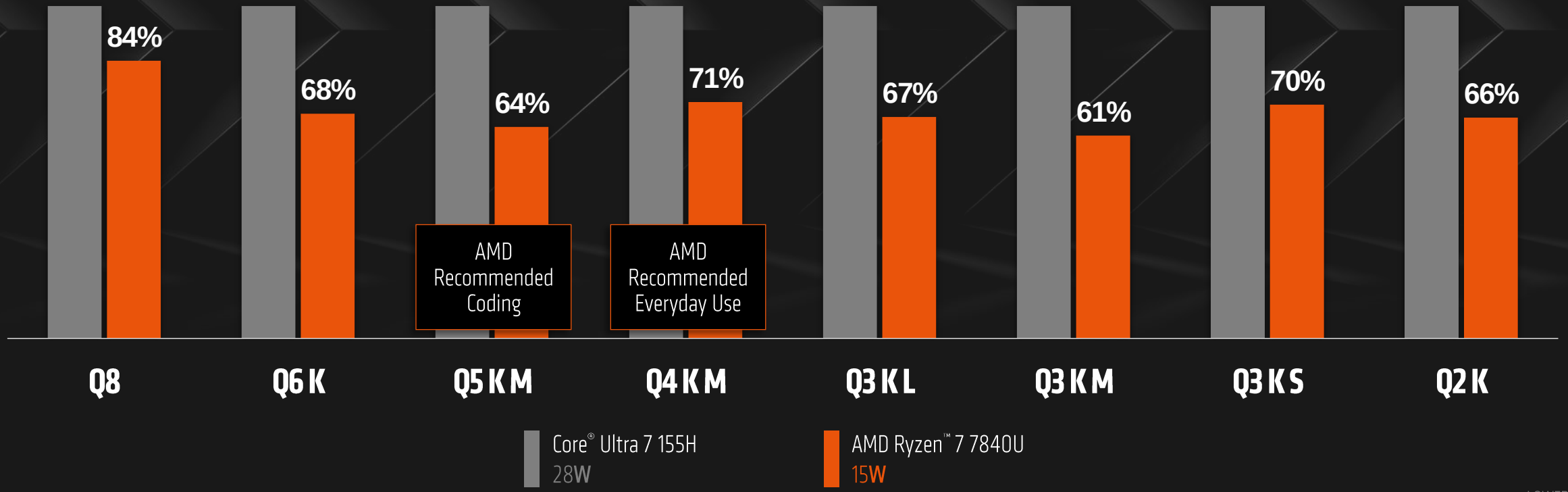
AMD
together we advance_AI

MISTRAL AI_

MISTRAL INSTRUCT 7B

TIME TO FIRST TOKEN WITH QUANTIZATION SCALING ON LM STUDIO

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI



LOWER IS BETTER

SEE ENDNOTE PHX-59

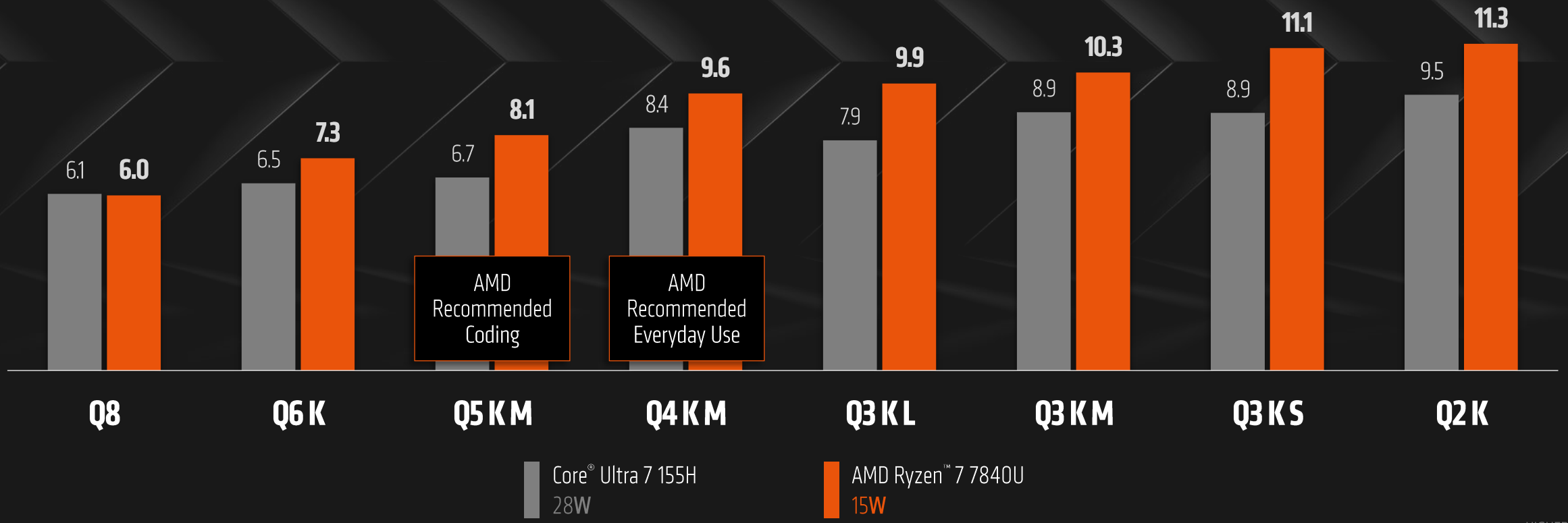
AMD
together we advance_AI

Llama 2

LLAMA 2 7B CHAT

TOKENS/S WITH QUANTIZATION SCALING ON LM STUDIO

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI



HIGHER IS BETTER

SEE ENDNOTE PHX-59

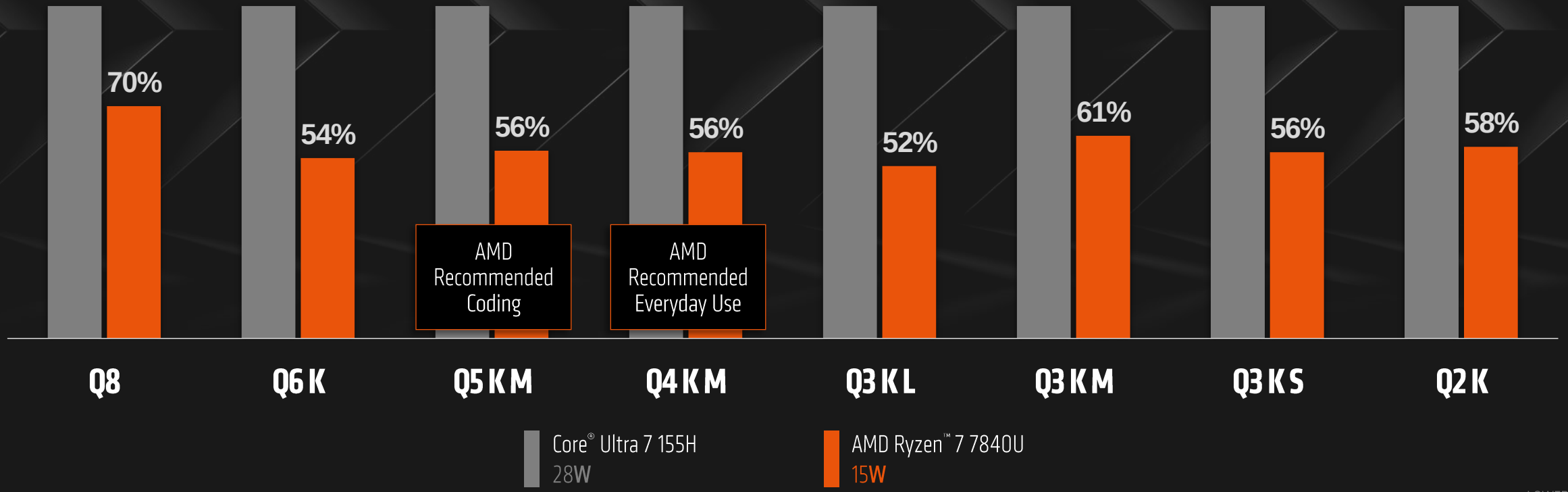
AMD
together we advance_AI

Llama 2

LLAMA 2 7B CHAT

TIME TO FIRST TOKEN WITH QUANTIZATION SCALING ON LM STUDIO

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI



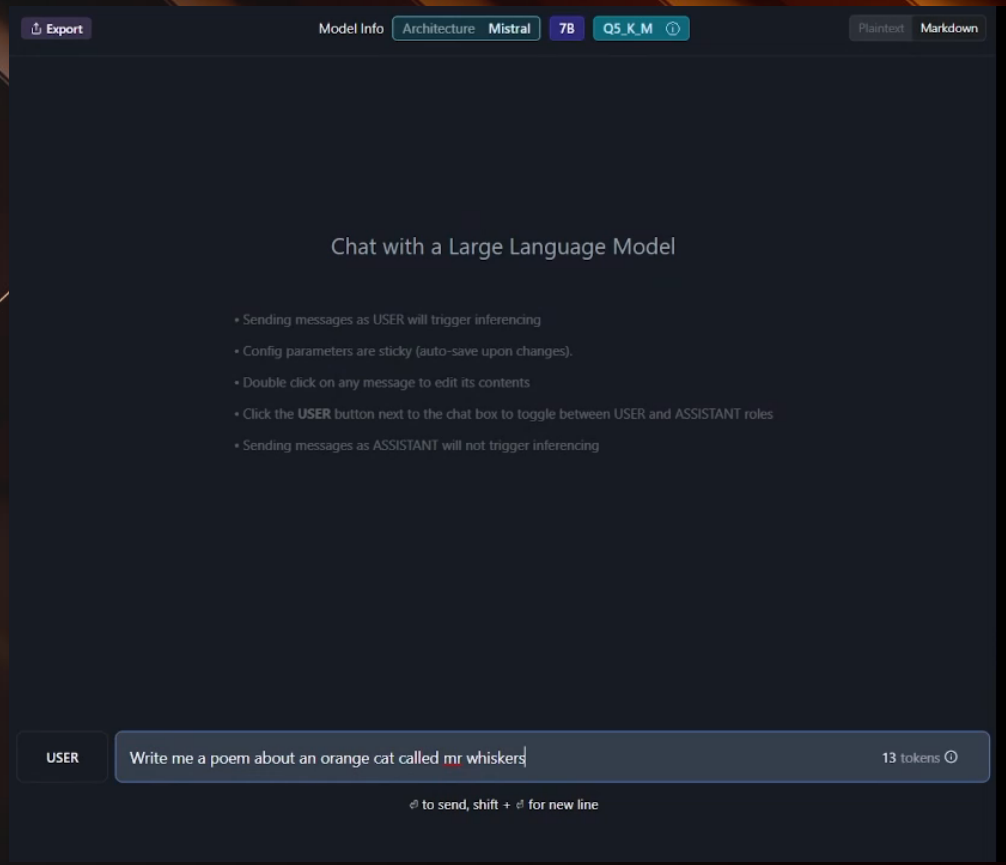
LOWER IS BETTER

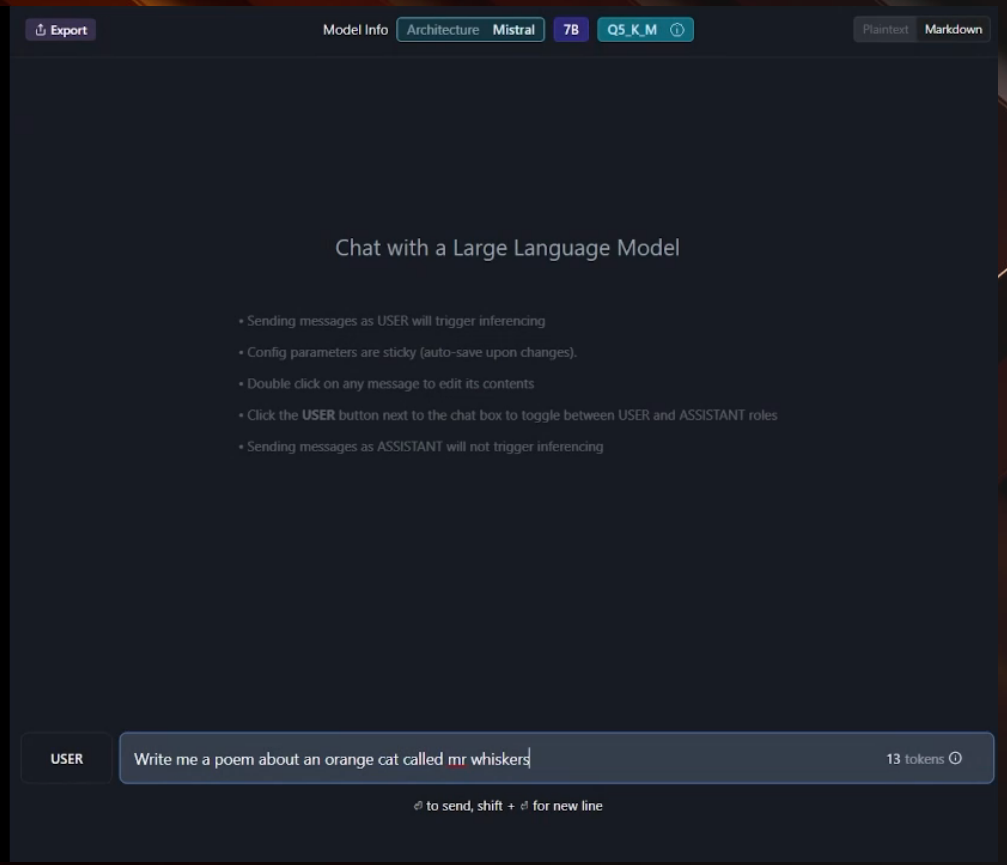
SEE ENDNOTE PHX-59

MISTRAL INSTRUCT 7B

SIDE BY SIDE DEMO

Intel® Core™ Ultra 7

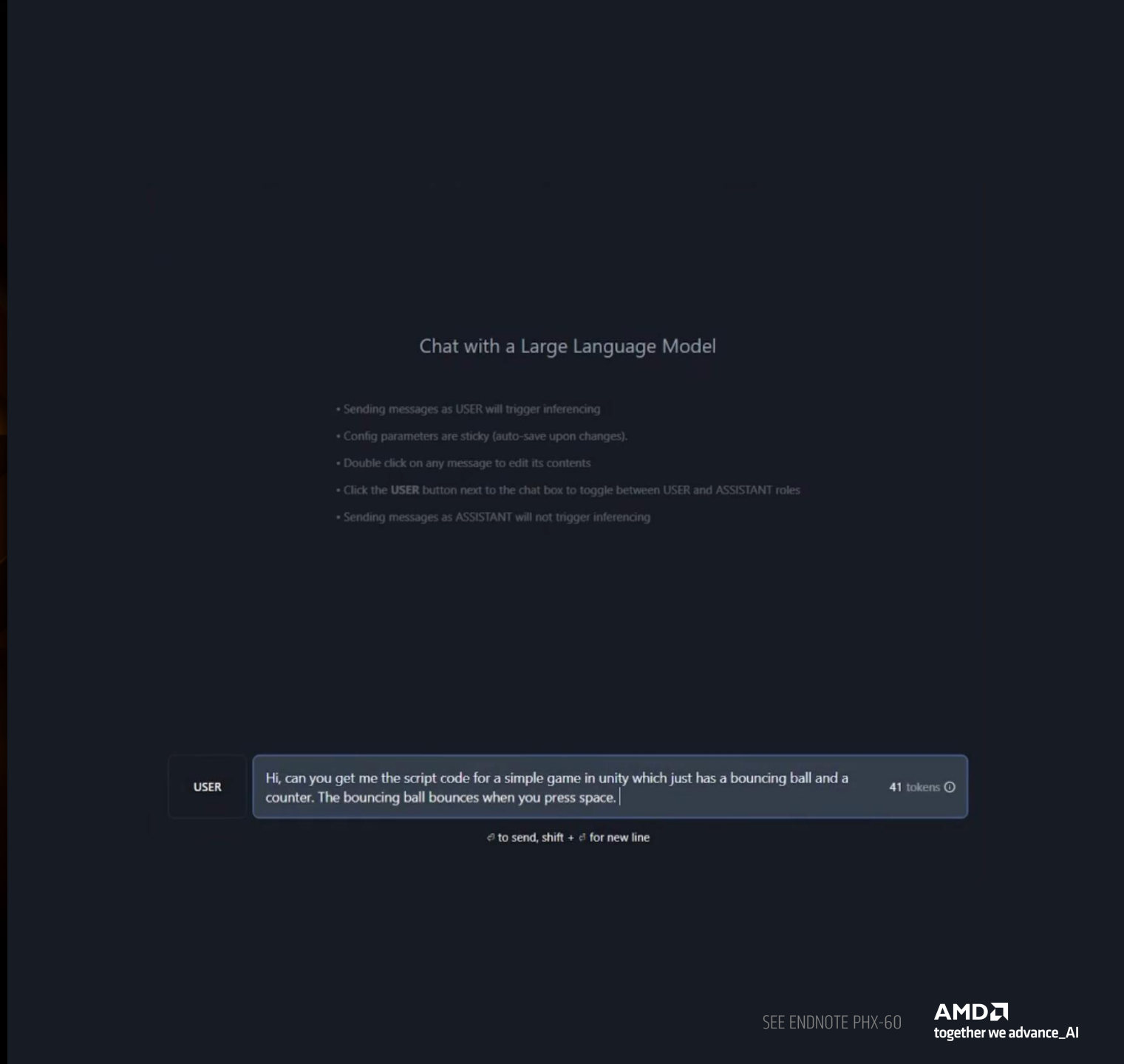




AMD RYZEN AI

SEE ENDNOTE PHX-61

Using *AMD Ryzen™ AI* for code generation on a local LLM

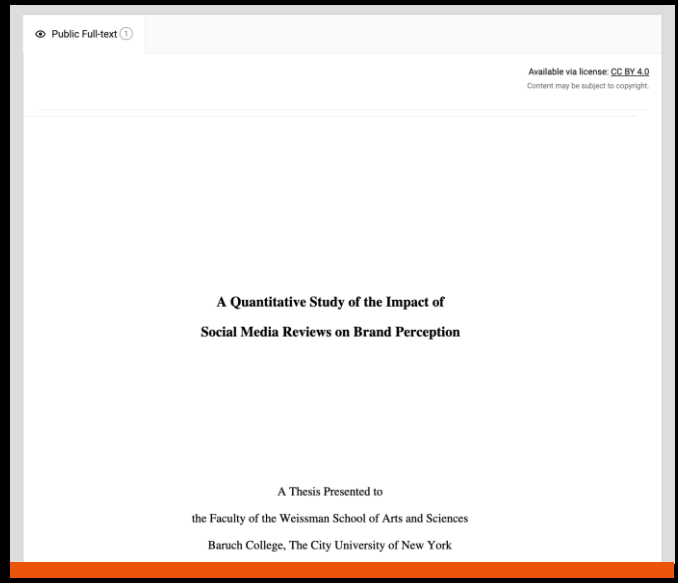


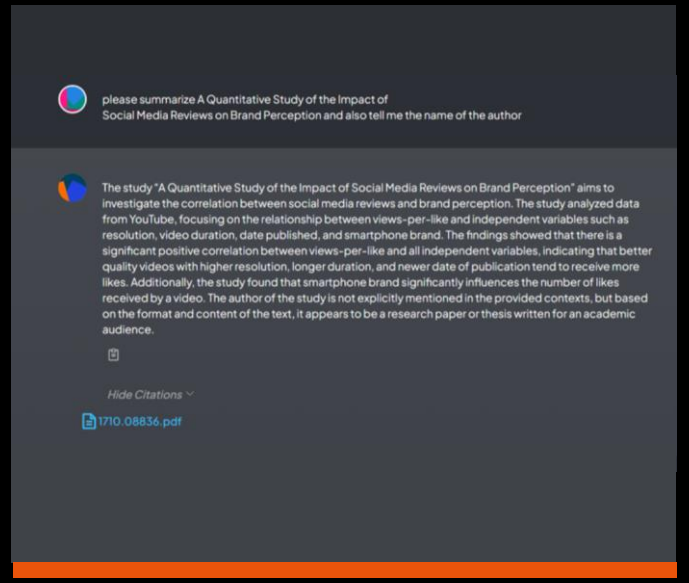
SEE ENDNOTE PHX-60

AMD
together we advance_AI

Improve the Quality of LLM responses by accessing local data sources

Retrieval Augmented Generation (RAG) running on AMD Ryzen™ AI





LM Studio Server > AnythingLLM Embedder > LanceDB Vector > Create Context Database

AMD
together we advance_AI

PLATFORM CHOICES FOR LOCAL AI

AMD RYZEN™ AI IS THE BEST SOLUTION

Intel Core Ultra Laptop	VS	AMD Ryzen™ AI Laptop
AI Capability		Leadership AI Performance
Higher Power (28W TDP)		Lower Power (15W TDP)
$999 SEP		**$899 SEP**
IPS 1920x1200 60Hz Screen		OLED IMAX Enhanced 2.8K 120Hz Screen
512 SSD		1 TB SSD

SEE ENDNOTE PHX-59

AMD
together we advance_AI

ENDNOTES

PHX-59: Testing as of Feb 2023 by AMD. Sustained performance average of multiple runs with specimen prompt ""Write me a story about an orange cat called mr whiskers"". All tests conducted on LM Studio 0.2.16. Performance may vary. Market price retrieved on 3/4/2023 (Amazon, US).
Phoenix: HP Pavilion Plus Laptop 14-ey0xxx, Ryzen 7 7840U 15W TDP, 16GB LPDDR5 6400, Windows 23H2 22631.3155, Adrenalin Driver 24.2.1
Meteor Lake: Acer Swift SFG14-72T, Intel Core Ultra 7 155H 28W TDP, 16GB LPDDR5 6400, Windows 23H2 22631.3155, Driver 31.0.101.5333.

PHX-60: Demo conducted by AMD on LM Studio 0.2.16 using Mistral Instruct 7b Q5 K M as of date presented. Performance may vary.
Demo made with prompt: "Hi, can you get me the script code for a simple game in unity which just has a bouncing ball and a counter. The bouncing ball bounces when you press space."
Phoenix: HP Pavilion Plus Laptop 14-ey0xxx, Ryzen 7 7840U 15W TDP, 16GB LPDDR5 6400, Windows 23H2 22631.3155, Adrenalin Driver 24.2.1

PHX-61: Demo conducted by AMD on LM Studio 0.2.16 using Mistral Instruct 7b Q5 K M. Performance may vary.
Demo made with prompt: "Write me a poem about an orange cat called mr whiskers."
Phoenix: HP Pavilion Plus Laptop 14-ey0xxx, Ryzen 7 7840U 15W TDP, 16GB LPDDR5 6400, Windows 23H2 22631.3155, Adrenalin Driver 24.2.1
Meteor Lake: Acer Swift SFG14-72T, Intel Core Ultra 7 155H 28W TDP, 16GB LPDDR5 6400, Windows 23H2 22631.3155, Driver 31.0.101.5333