

AMD 
RYZEN AI

CONSUMER AI PERFORMANCE
WITH LARGE LANGUAGE MODELS (LLMs)

MARCH 2024

AMD 
together we advance AI

RUNNING AN LLM ON YOUR PROCESSOR

WITH LM STUDIO



100 Million

users within 2 months
on ChatGPT

LOCAL  PRIVACY

NO SUBSCRIPTION FEES

NO WI-FI NEEDED



Popular LLM



Comparable to
GPT 3.5/4 models



Used for coding
with markdown

RAG

Local context
for LLMS

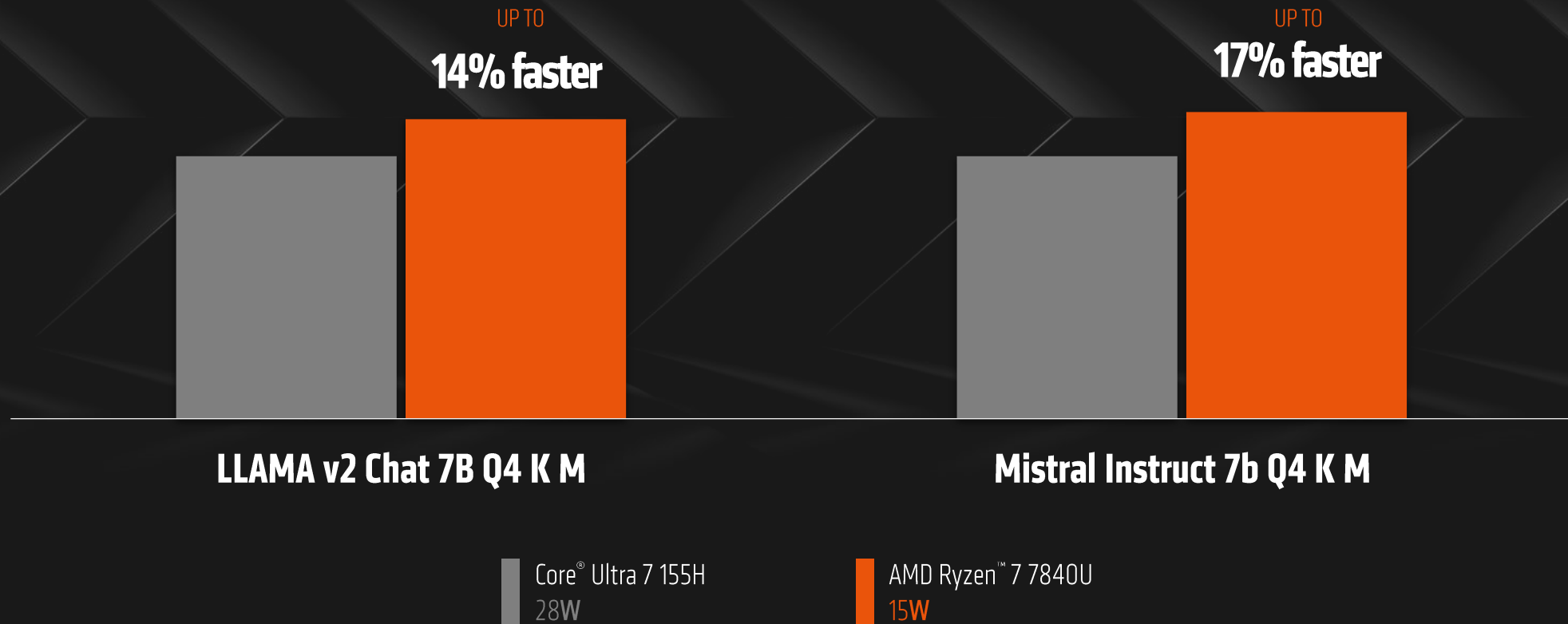
GET YOUR LLM UP *AND* *RUNNING*

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



LEADING LLM PERFORMANCE ON AMD RYZEN™ AI

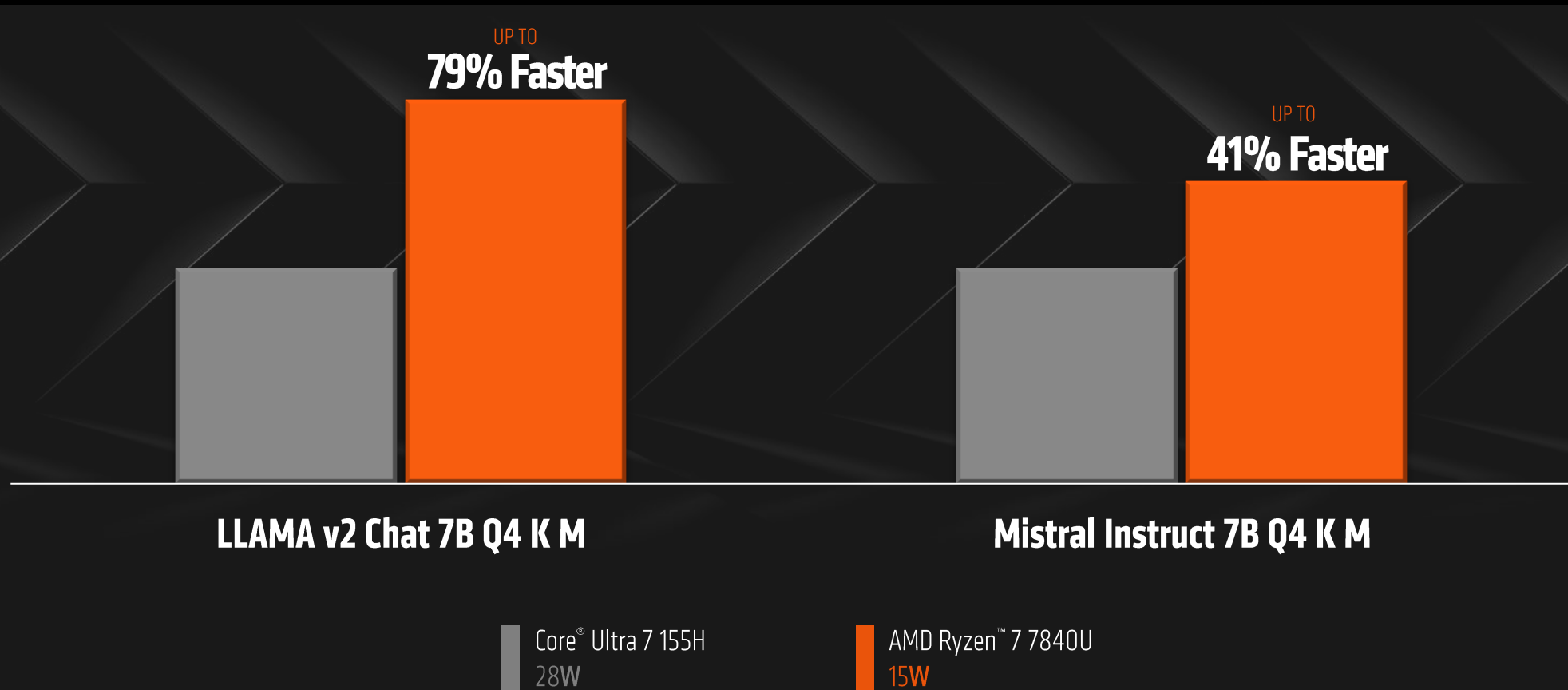
TOKENS PER SECOND FOR SAMPLE PROMPT ON LM STUDIO



SEE ENDNOTE PHX-59

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI

TIME TO FIRST TOKEN FOR SAMPLE PROMPT ON LM STUDIO



HIGHER IS BETTER

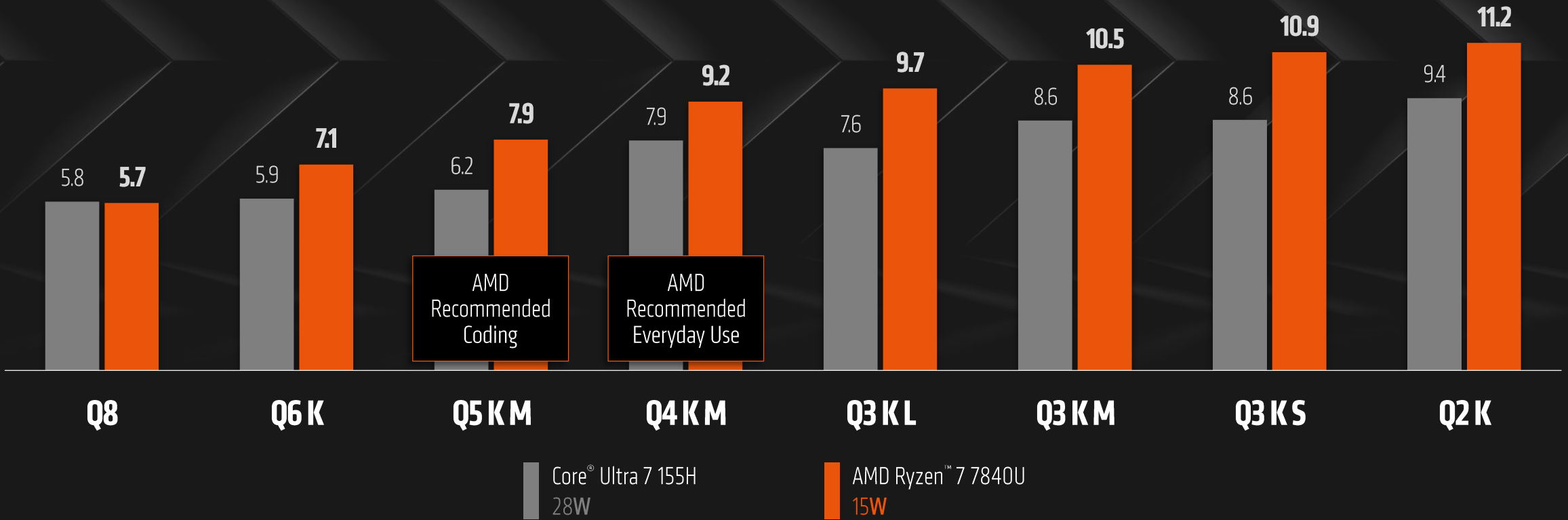
SEE ENDNOTE PHX-59



MISTRAL INSTRUCT 7B

TOKENS/S WITH QUANTIZATION SCALING ON LM STUDIO

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI



HIGHER IS BETTER

SEE ENDNOTE PHX-59

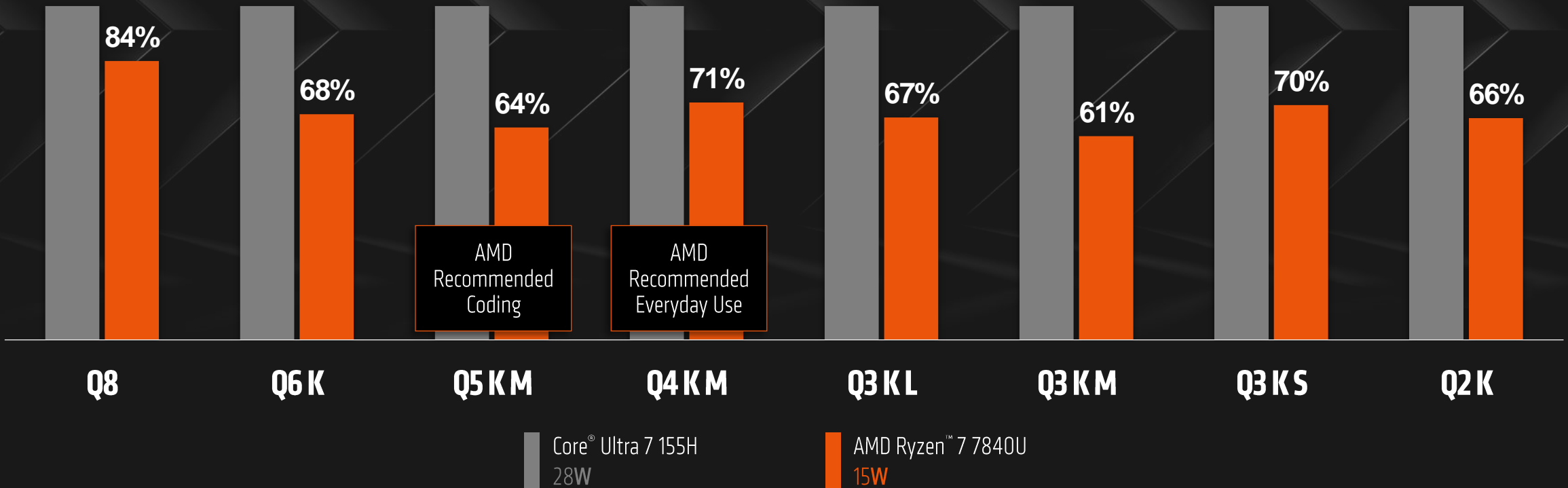




MISTRAL INSTRUCT 7B

TIME TO FIRST TOKEN WITH QUANTIZATION SCALING ON LM STUDIO

LEADING LLM PERFORMANCE ON AMD RYZEN™ AI



LOWER IS BETTER

SEE ENDNOTE PHX-59

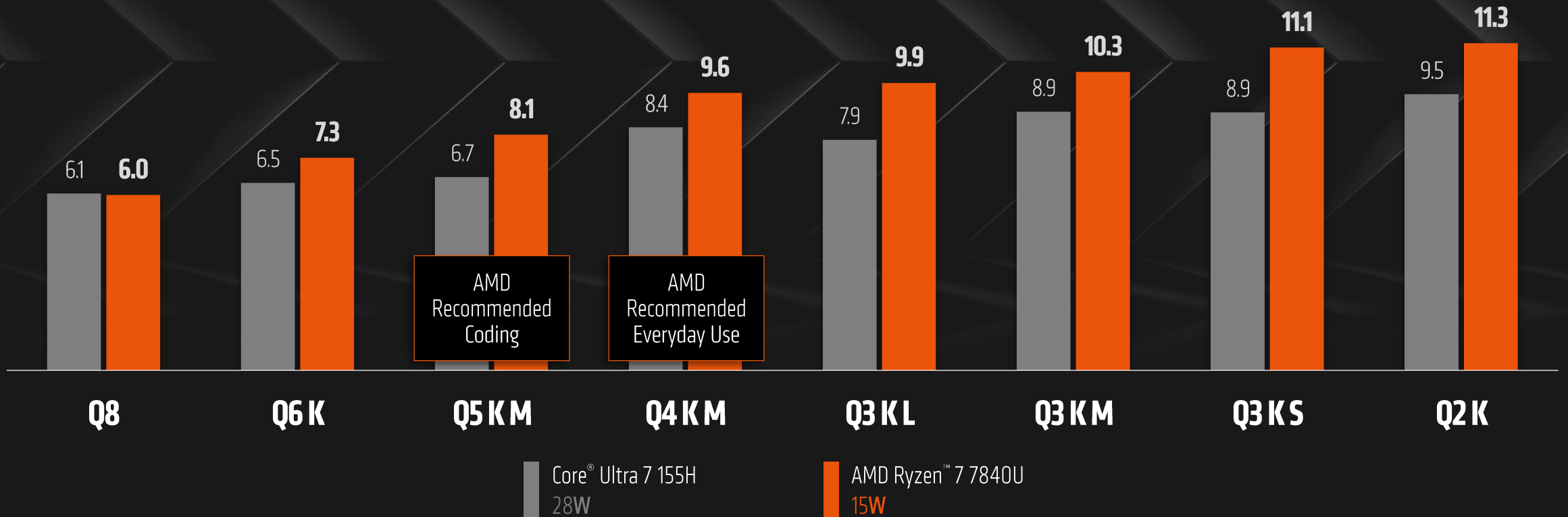


Llama 2

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

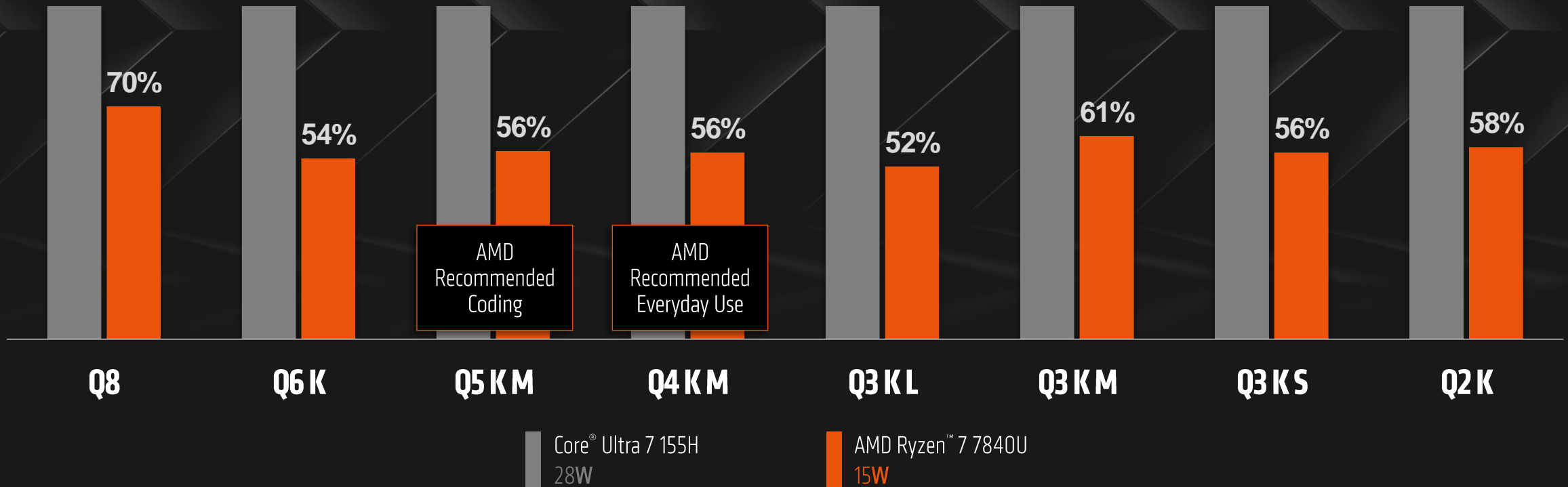
AMD
together we advance AI

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

AMD
together we advance AI



MISTRAL INSTRUCT 7B

SIDE BY SIDE DEMO

Export Model Info Architecture Mistral 7B Q5_K_M Plaintext Markdown

Chat with a Large Language Model

- Sending messages as USER will trigger inferencing
- Config parameters are sticky (auto-save upon changes).
- Double click on any message to edit its contents
- Click the USER button next to the chat box to toggle between USER and ASSISTANT roles
- Sending messages as ASSISTANT will not trigger inferencing

USER Write me a poem about an orange cat called mr whiskers 13 tokens

to send, shift + for new line

Export Model Info Architecture Mistral 7B Q5_K_M Plaintext Markdown

Chat with a Large Language Model

- Sending messages as USER will trigger inferencing
- Config parameters are sticky (auto-save upon changes).
- Double click on any message to edit its contents
- Click the USER button next to the chat box to toggle between USER and ASSISTANT roles
- Sending messages as ASSISTANT will not trigger inferencing

USER Write me a poem about an orange cat called mr whiskers 13 tokens

to send, shift + for new line

Intel® Core™
Ultra 7

AMD
RYZEN AI

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

Chat with a Large Language Model

- Sending messages as USER will trigger inferencing
- Config parameters are sticky (auto-save upon changes).
- Double click on any message to edit its contents
- Click the **USER** button next to the chat box to toggle between USER and ASSISTANT roles
- Sending messages as ASSISTANT will not trigger inferencing

USER

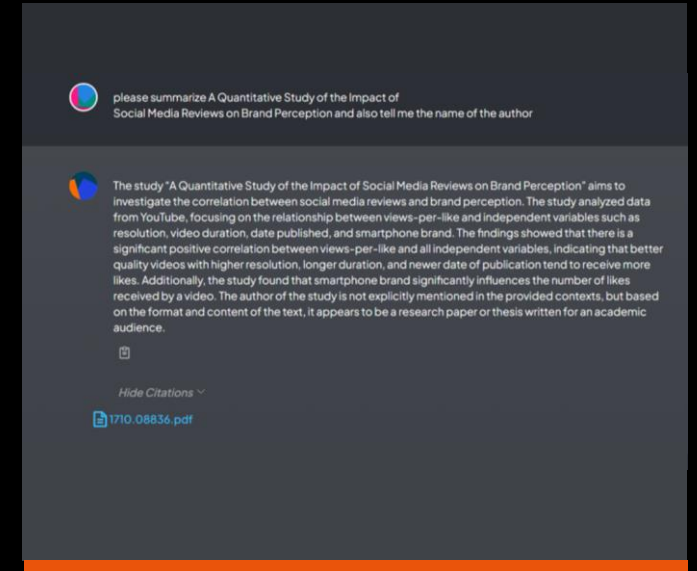
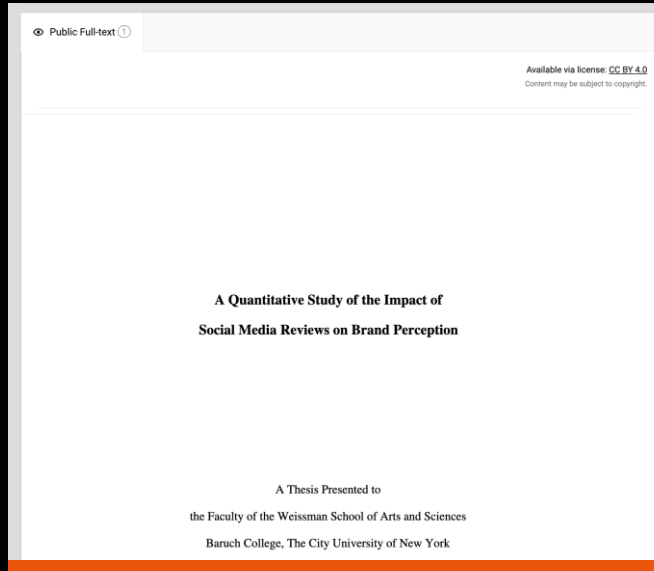
Hi, can you get me the script code for a simple game in unity which just has a bouncing ball and a counter. The bouncing ball bounces when you press space.

41 tokens ⓘ

↵ to send, shift + ↵ for new line

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

PLATFORM CHOICES FOR LOCAL AI

AMD RYZEN™ AI IS THE BEST SOLUTION



Intel Core Ultra Laptop

AI Capability

Higher Power
(28W TDP)

\$999 SEP

IPS 1920x1200
60Hz Screen

512 SSD

VS

AMD Ryzen™ AI Laptop

Leadership AI Performance

Lower Power
(15W TDP)

\$899 SEP

OLED IMAX Enhanced
2.8K 120Hz Screen

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

AMD

together we advance_AI

GENERAL DISCLAIMER: THE INFORMATION CONTAINED HEREIN IS FOR INFORMATIONAL PURPOSES ONLY AND IS SUBJECT TO CHANGE WITHOUT NOTICE. WHILE EVERY PRECAUTION HAS BEEN TAKEN IN THE PREPARATION OF THIS DOCUMENT, IT MAY CONTAIN TECHNICAL INACCURACIES, OMISSIONS AND TYPOGRAPHICAL ERRORS, AND AMD IS UNDER NO OBLIGATION TO UPDATE OR OTHERWISE CORRECT THIS INFORMATION. ADVANCED MICRO DEVICES, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS DOCUMENT, AND ASSUMES NO LIABILITY OF ANY KIND, INCLUDING THE IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY OR FITNESS FOR PARTICULAR PURPOSES, WITH RESPECT TO THE OPERATION OR USE OF AMD HARDWARE, SOFTWARE OR OTHER PRODUCTS DESCRIBED HEREIN. NO LICENSE, INCLUDING IMPLIED OR ARISING BY ESTOPPEL, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. TERMS AND LIMITATIONS APPLICABLE TO THE PURCHASE OR USE OF AMD'S PRODUCTS ARE AS SET FORTH IN A SIGNED AGREEMENT BETWEEN THE PARTIES OR IN AMD'S STANDARD TERMS AND CONDITIONS OF SALE. GD-18

© 2024 ADVANCED MICRO DEVICES, INC. ALL RIGHTS RESERVED. AMD, THE AMD ARROW LOGO, RADEON, RYZEN, AND COMBINATIONS THEREOF ARE TRADEMARKS OF ADVANCED MICRO DEVICES, INC. OTHER PRODUCT NAMES CONTAINED HEREIN ARE FOR IDENTIFICATION PURPOSES ONLY AND MAY BE TRADEMARKS OF THEIR RESPECTIVE OWNERS.