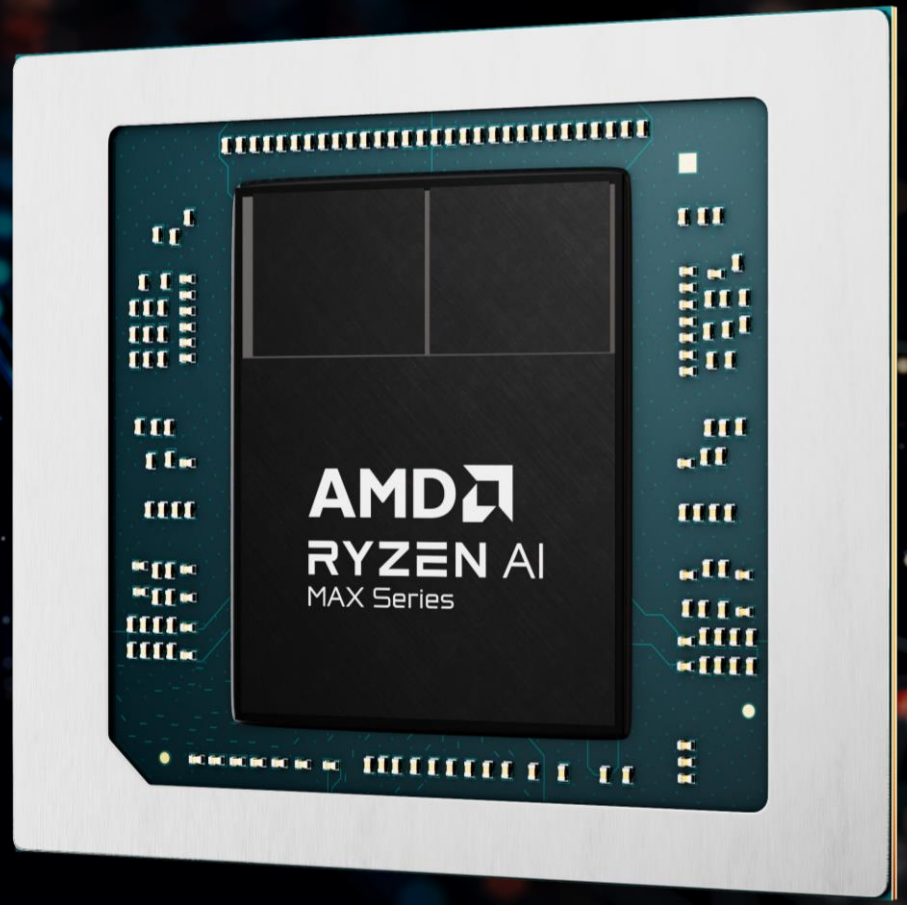
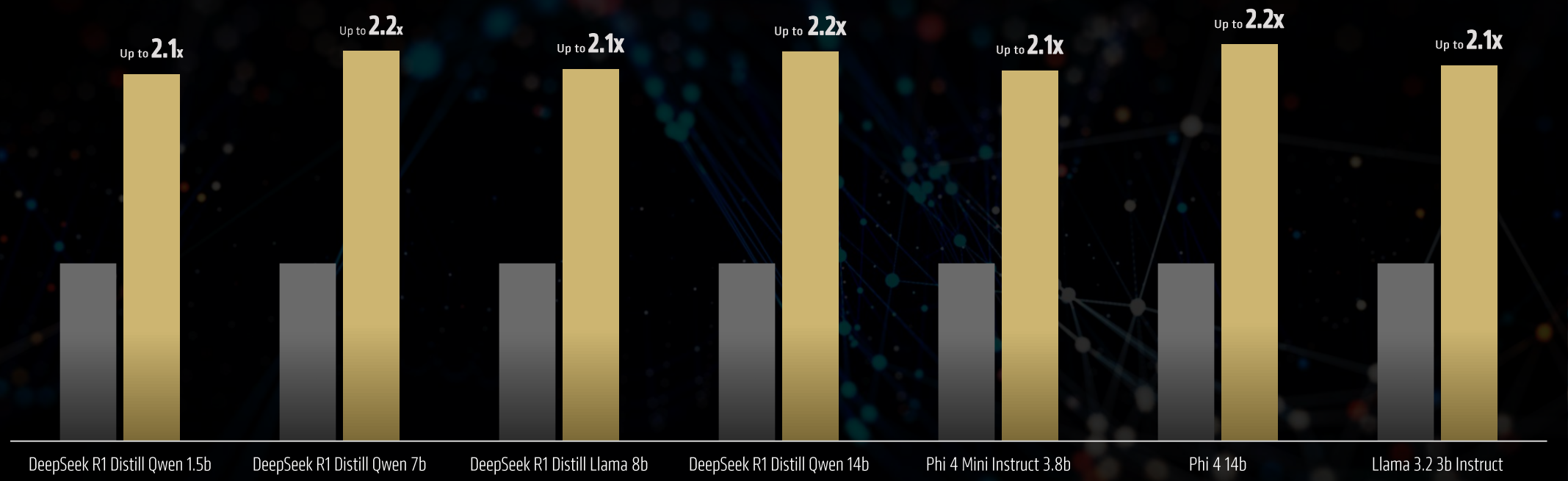
AMD *RYZEN™ AI* MAX+ 395
AI PERFORMANCE



MOST POWERFUL X86 PROCESSOR FOR LLMs

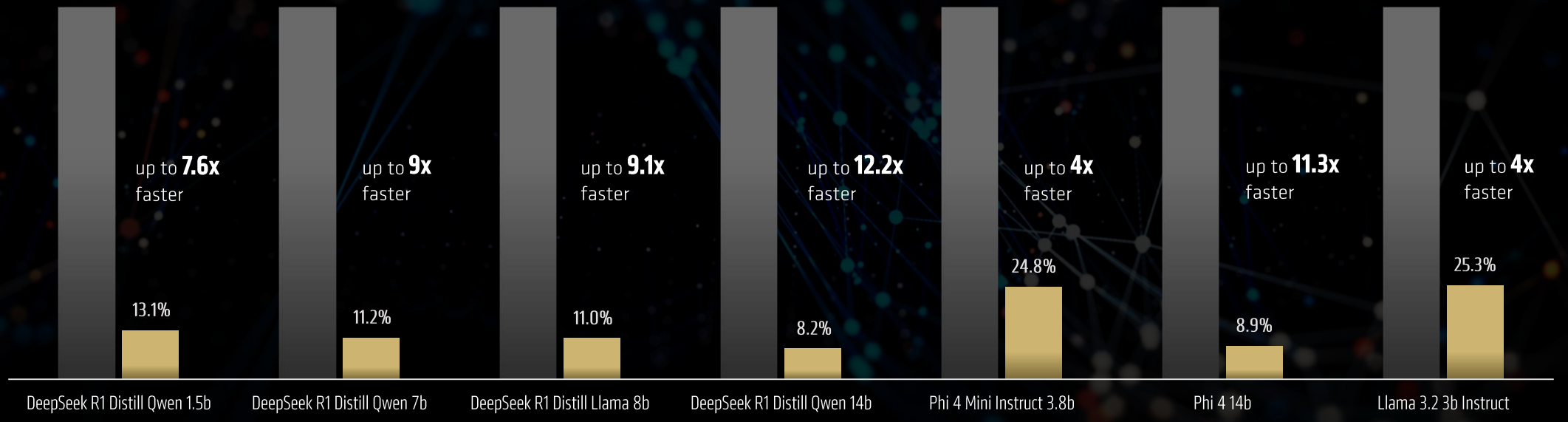
LM STUDIO 0.3.11 – VARIOUS LLMS (< 16GB MODEL SIZE) – *TOKENS PER SECOND*



Testing as of March 2025 by AMD. All tests conducted on LM Studio 0.3.11. Llama.cpp runtime 1.18. Tokens/s and time to first token: Sustained performance average of multiple runs with specimen prompt "How long would it take for a ball dropped from10 meter height to hit the ground?". Models tested: DeepSeek R1 Distill Qwen 1.5b Q4 K M, DeepSeek R1 Distill Qwen 7b Q4 K M, DeepSeek R1 Distill Qwen 8b Q4 K M, DeepSeek R1 Distill Qwen 14b Q4 K M, Phi 4 Mini Instruct 3.8b, Phi 4 Q4 K M, Llama 3.2 3b Instruct. AMD Ryzen™ AI MAX+ 395 on an ASUS ROG Flow Z13 with 64GB 8000 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.3.1 WHQL. VGM = 32GB Intel Core Ultra 7 258V on an HP Zenbook S14 with 32GB 8533 MT/s memory, Windows 11 Pro 24H2 and Intel Graphics Driver 32.0.101.6559. Performance may vary. SHO-26.

UP TO 12.2X FASTER THAN THE COMPETITION IN TEXT MODELS

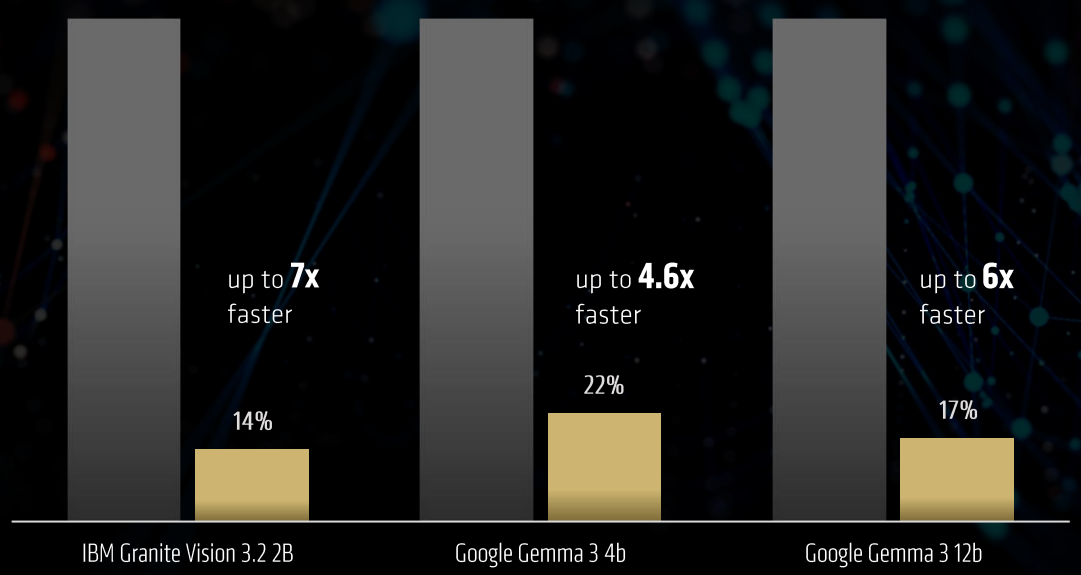
LM STUDIO 0.3.11 – VARIOUS LLMS (< 16GB MODEL SIZE) – *TIME TO FIRST TOKEN*



Testing as of March 2025 by AMD. All tests conducted on LM Studio 0.3.11. Llama.cpp runtime 1.18. Tokens/s and time to first token: Sustained performance average of multiple runs with specimen prompt "How long would it take for a ball dropped from 10 meter height to hit the ground?". Models tested: DeepSeek R1 Distill Qwen 1.5b Q4 K M, DeepSeek R1 Distill Qwen 7b Q4 K M, DeepSeek R1 Distill Qwen 8b Q4 K M, DeepSeek R1 Distill Qwen 14b Q4 K M, Phi 4 Mini Instruct 3.8b, Phi 4 Q4 K M, Llama 3.2 3b Instruct. AMD Ryzen™ AI MAX+ 395 on an ASUS ROG Flow Z13 with 64GB 8000 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.3.1 WHQL. VGM = 32GB Intel Core Ultra 7 258V on an HP Zenbook S14 with 32GB 8533 MT/s memory, Windows 11 Pro 24H2 and Intel Graphics Driver 32.0.101.6559. Performance may vary. SHO-26.

UP TO 6X FASTER IN SOTA VISION MODELS

LM STUDIO 0.3.13 – VISION MODELS - *TIME TO FIRST TOKEN*



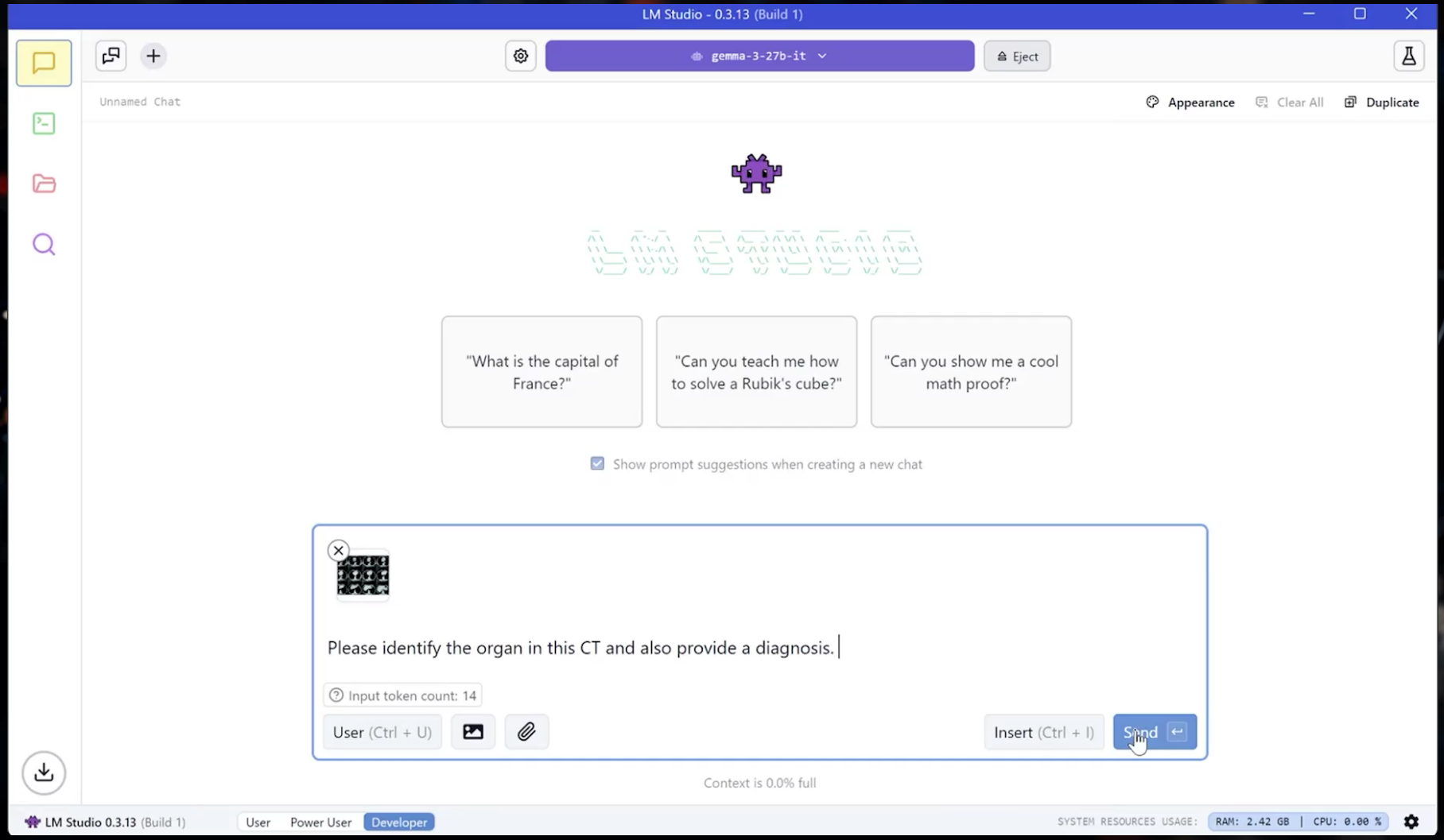
Query: Describe this image

ASUS Zenbook S14 32 GB - Intel Core Ultra 7 258V (Arc 140V)

ASUS ROG Flow Z13 64 GB- AMD Ryzen™ AI Max+ 395 (AMD Radeon™ 8060S)

Testing as of March 2025 by AMD. All tests conducted on LM Studio 0.3.13. Llama.cpp runtime 1.19.2. Time to first token using prompt "Describe this image" and picture taken from: www.loc.gov/item/2013648266/ Models tested: IBM Granite Vision 3.2 2b, Google Gemma 3 4b, Google Gemma 3 12b and Google Gemma 3 27b AMD Ryzen™ AI MAX+ 395 on an ASUS ROG Flow Z13 with 64GB 8000 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.3.1 WHQL. VGM = 32GB Intel Core Ultra 7 258V on an HP Zenbook S14 with 32GB 8533 MT/s memory, Windows 11 Pro 24H2 and Intel Graphics Driver 32.0.101.6559. Links to third party sites are provided for convenience and unless explicitly stated, AMD is not responsible for the contents of such linked sites and no endorsement is implied. Performance may vary. SHO-27.

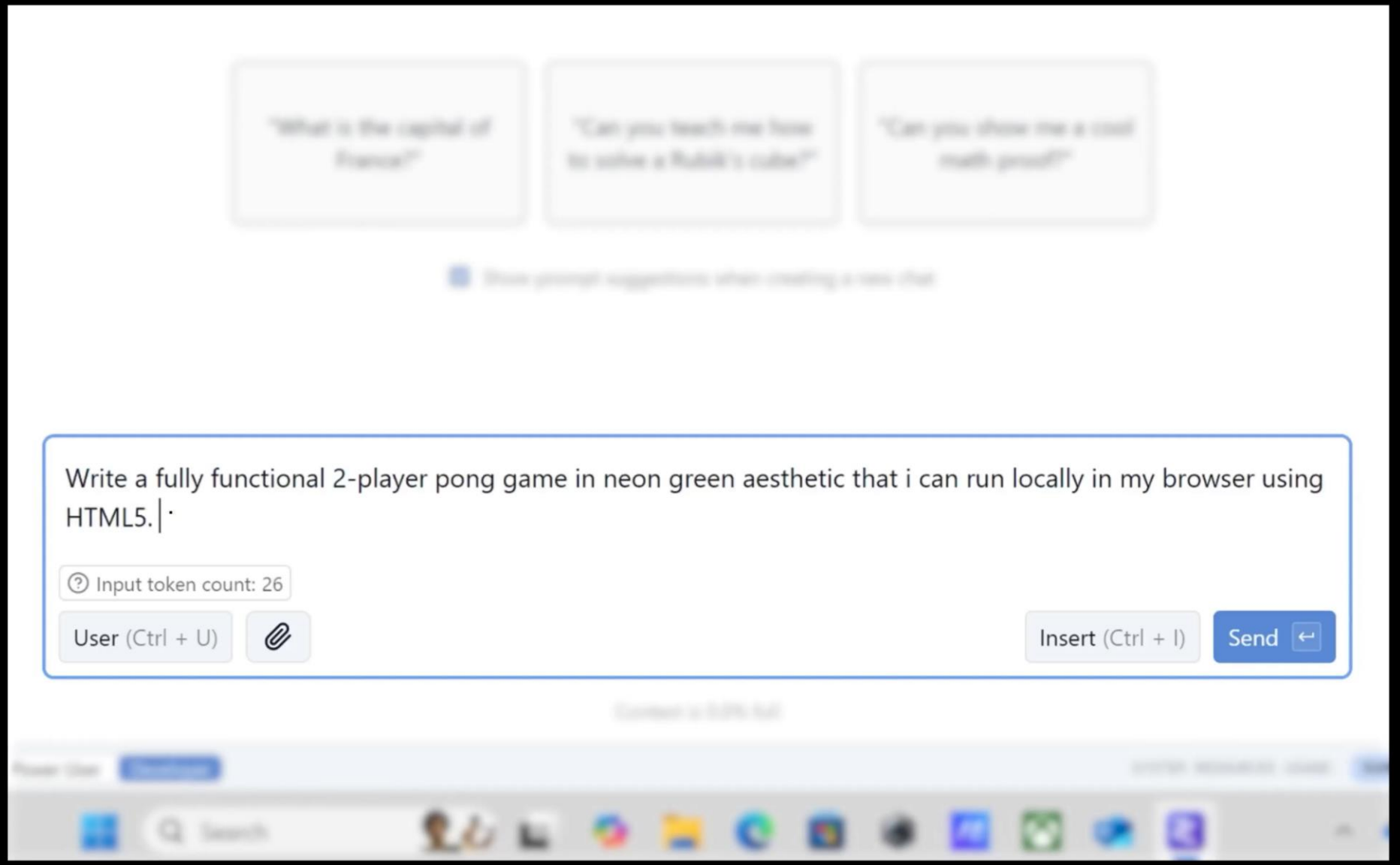
FASTEST X86 PROCESSOR FOR GOOGLE GEMMA 3 27b VISION MODEL



See Endnote: SH0-27

Pong, Brick Breaker and Tetris-like Games Created On Strix Halo 64GB

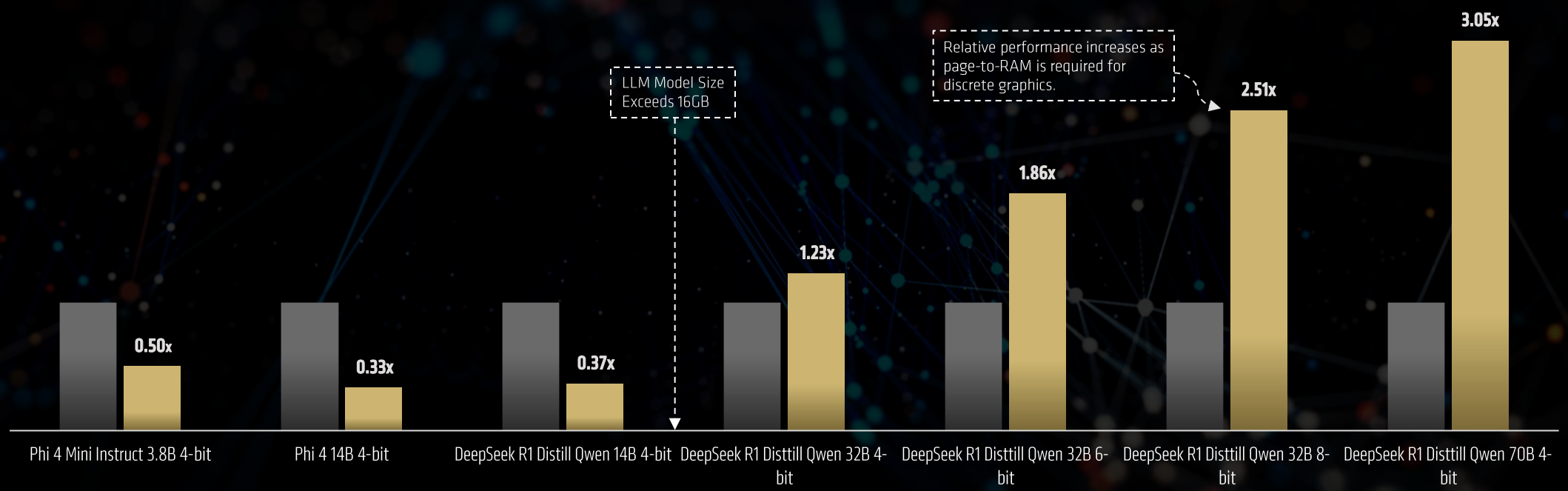
Recreating a gaming classic took 5 minutes with no coding knowledge



See Endnote: SH0-26

AMD RYZEN™ AI 9 MAX 395+ PROCESSOR 128GB

LM STUDIO 0.3.11 – *TOKENS PER SECOND* – LARGE LLMS AND QUANTIZATIONS REQUIRE LARGE VRAM



Testing as of March 2025 by AMD. All tests conducted on LM Studio 0.3.11. Llama.cpp runtime 1.18. Tokens/s using prompt "How long would it take for a ball dropped from10 meter height to hit the ground?" Models tested: Phi 4 Mini Instruct Q4 K M, Phi 4 Q4 K M, DeepSeek R1 Distill Qwen 14b Q4 K M, DeepSeek R1 Distill Qwen 32b Q4 K M, DeepSeek R1 Distill Qwen 32b Q6, DeepSeek R1 Distill Qwen 32b Q8 and DeepSeek R1 Distill Llama 70b Q4 K M AMD Ryzen™ AI MAX+ 395 CRB 55W with 128GB 8000 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.3.1 WHQL. NVIDIA GeForce RTX 5080 16GB with a Ryzen 7 9800 X3D and 32GB 6000 MT/s memory, Windows 11 Pro 24H2 and GeForce 572.47 drivers. Performance may vary. SHO-28.