Inkling 2 Mini on an RTX 5090

What Squid5090 gains from speculative decoding, how it handles long prompts, and where a batching server helps.



Original AI-generated hardware illustration. The Squid5090 and SGLang comparisons used one RTX 5090 per engine. The FlashNext quality baseline used four.

372.1
Output tokens/s in one Squid chat at temperature 1

516,092
Input tokens with all three target codes recovered

18 / 21
General battery score with normal Inkling reasoning

The benchmark was run on 26 November 2026 UTC. It covers single-stream output, a four-request burst, three long-context trials, an eight-document extraction replay and a 21-case quality battery.

Output speed and concurrency

Squid5090 keeps the checkpoint's mixed-precision expert weights, while the measured FP8 arm uses FP8 dense attention projections, output head and KV cache. CUDA graphs cut host launch overhead. DSpark can verify several positions per pass.

Context
Existing KV cache
→
Draft
Up to seven proposals
→
Verify
One target-model pass
→
Emit
Accepted tokens

Single-stream chat at temperature 1



Output tokens per second. Two runs per configuration, up to 384 output tokens. Rates can include reasoning tokens. Both engines used a 32,768-token setting, with FP8 KV in Squid and BF16 KV in SGLang.

Measurement	Squid5090	SGLang
Cold 18,857-token prompt, time to first token	1.225 s	1.291 s
Warm repeat, time to first token	0.156 s	0.051 s
Four-request burst, worst time to first token	3.318 s	0.187 s
Four-request burst, aggregate output	355.6 tokens/s	469.0 tokens/s

Squid served the four requests in sequence. SGLang batched them. The concurrency figures are one measured burst, with separate cards and the specific engine configurations shown above.

Build and measurement

Linux, NVIDIA driver 615.71.09, CUDA 13.4.92 and Zig 0.17.1 for Squid. The client used usage-reported tokens and timestamps on the first and last output-bearing SSE deltas. See the machine-readable results for prompt lengths and completion status.

Long prompts and answer quality

Configured context	Request and outcome
524,288 · FP8/Q4	516,090 input tokens with reasoning on. The 100-token output budget went to reasoning. No target code was returned.
524,288 · FP8/Q4	516,092 input tokens, 3/3 exact codes at roughly 10%, 50% and 90% depth. Thinking disabled. First token at 118.47 s.
1,048,576 · NVFP4/Q4	Server reached readiness. The request targeted about 1.03 million tokens and did not stream a response within the 405 s client deadline.

> A successful load establishes that the configuration fits. Completed requests establish usable context. For coding sessions, cold prefill and warm continuation both matter.

The 21-case battery

Mode	Passed	Sum of request time
Inkling with normal reasoning · Squid	18 / 21	21.6 s
Inkling with thinking disabled · Squid	15 / 21	6.3 s
FlashNext Hybrid · vLLM TP4	17 / 21	60.1 s

All runs used temperature 0 and the same test cases. The FlashNext row is a different model and deployment. One fast-mode answer used **3,901** where the scorer expected the literal **3901**. The discount calculation, quicksort and tool-call case show genuine behavior differences.

Eight-document extraction replay

With `enable_thinking: false`, Squid returned eight responses that parsed as JSON objects in a mean of 1.36 seconds per document. Setting `reasoning_effort: none` alone produced six such responses at 4.34 seconds per document. Field disagreements remained after the JSON check. General code and tool tasks benefit from retaining normal reasoning.

Sources and credits

Measurements by senn-tech, 26 November 2026 UTC. Public sources checked 27 November 2026. The benchmark data contains per-run metrics and battery case verdicts. The report uses aggregate importer observations.

Tensorfish. Squid5090 engineering article and tested repository revision.
`124afb85116c89b375925606bd129ba7fc4ec1b7`

Upstream Squid contributions. The project NOTICE credits satellitedown's ingine work for major kernels and optimizations. It also credits Atomic Engine, Sybil Solutions, stb and libwebp for distinct contributions.

Thinking Machines Lab. Inkling 2 Mini model card and checkpoint.
`c9d028302cd0f035fdf8e7e91f089a4ce94d2195`

DeepSeek. DeepSpec repository and DSpark.

SGLang contributors. Inkling integration and official recipe.
`Integration Git 48177212585eeaabca4db7e434c00a9c8a7ae604`
`Tested image sha256:efac853b46811ccb760545b7b29e7119bdf907537ccf6e863fac3011fe8b98de`

vLLM and NVIDIA. vLLM 0.32.0 release, the base of the local FlashNext lane, and the GeForce RTX 5090 specifications.

THINKING MACHINES

Thinking Machines Lab via Wikimedia Commons. Public domain per file page.

NVIDIA mark via Simple Icons v13. Trademark belongs to NVIDIA.

Cover illustration generated for senn-tech with OpenAI Image Generation. Charts and diagrams use the measurements and cited source code.