The KV-Cache Patent White Space in Edge AI

Inflexion Report excerpt · AI Accelerator Chip Architecture · April 2026

Executive Briefing

The KV-cache management problem is the single largest unsolved bottleneck in deploying large language models on edge hardware. Yet the patent landscape for hardware-level KV-cache solutions is almost empty. The window to file defensible IP is measured in months, not years — major players are watching this space and the first mover advantage is substantial.

White Space Recommendation #1 — Priority Scorecard

URGENCY	PRIORITY SCORE	INVESTMENT RANGE	FILING WINDOW
6 MONTHS	**9.5 / 10**	**£150K – £400K**	**NOW – Q4 2026**
before major players file defensively	for 5-8 provisional applications	for 5–8 provisional applications	before Apple/Qualcomm/ Samsung file defensively

Four High-Value Claim Areas

#	Claim Area	Why It Matters	Landscape
A	Hardware KV-cache eviction policies	Determines which tokens stay in fast SRAM vs spill to DRAM — 10× latency gap	Near-empty
B	NVM persistence for KV-cache	Non-volatile storage of KV state enables session continuity without full recompute	Near-empty
C	Mixed-precision KV-cache (FP8/INT4)	4× compression of KV state without accuracy loss — critical for 1–4 GB edge DRAM budgets	Sparse
D	Speculative decoding hardware	Draft model + verification model running concurrently — 2–4× inference speedup	Sparse

■ Who Should Act

Fabless chip startups with valuations below $500M — large enough to execute a patent programme, small enough that a defensive IP portfolio creates disproportionate competitive moat. Particularly relevant for companies building edge inference ASICs for on-device LLM serving.

■ Risk If Ignored

Within 18 months, Apple, Qualcomm, Samsung, and Google are expected to file defensive portfolios covering hardware KV-cache management. Once filed, the white space closes permanently and smaller players face either licensing costs or design-around constraints.

The full AI Accelerator Inflexion Report covers 50+ patents, photonic tensor core analysis, spiking neural network convergence signals, and a prioritised R&D investment roadmap across 6 white space areas.

Full Inflexion Report: £2,500 · 50–70 pages · 2–3 week delivery