Ollama Full-Pass Analysis — Status & Methodology

Quran Ollama Full-Pass — Status Report

Generated: 2026-03-05 **MC Task:** #1949 (previous failed), this run: test pass **Status:** TEST CHUNKS COMPLETE — 3/35 chunks verified, production run ready

---

Model Used

qwen2.5-coder:32b (19GB, Ollama local, Mac Studio M3 Ultra 96GB RAM)

Rationale:

- Largest available reasoning model in the Ollama lineup on this machine
- 32B parameters gives substantially better analytical depth than llama3.1:8b
- M3 Ultra's 96GB unified memory easily accommodates 19GB model + up to 32K context
- qwen2.5-coder is an instruction-following model with strong analytical capability despite the "coder" label — verified via test chunks

Other models available (not used):

- `llama3.1:8b` — fallback, 4.9GB, faster but lower quality reasoning
- `alaiml-task-v1`, `alaiml-email-v1`, `alaiml-tender-v1` — custom fine-tunes, not suitable for Quran analysis
- `llama-guard3:8b` — safety classification model, not suitable
- `bge-m3:latest` — embedding model only
- `nomic-embed-text:latest` — embedding model only

Chunking Strategy

Algorithm: Ayah-budget-aware greedy grouping with special handling for very large surahs.

Target: ~180 ayahs per chunk **Total chunks:** 35 (covering all 114 surahs, 6,236 ayahs)

Rules:

1. Surahs with ≥180 ayahs get their own dedicated chunk (prevents context overflow from a single massive surah)
2. Remaining surahs are greedily grouped: add next surah if total stays ≤180 ayahs AND group already has ≥2 surahs
3. This creates semantically coherent groupings that stay within Ollama's processing ability

Why this beats the previous approach (MC Task #1949):

- Previous attempt used the Claude agent runner which hit max turns — each surah was one API call to a Claude agent session, causing turn exhaustion at ~50+ surahs
- New approach: standalone Node.js script calls Ollama directly via HTTP — no turn limit, no session state, pure function calls
- Resume capability via file existence check — if chunk file exists, skip. Can be interrupted and restarted at any point

Chunk plan (all 35):

Chunk	Ayahs	Surahs
01	7	1: Al-Faatiha
02	286	2: Al-Baqara
03	200	3: Aal-i-Imraan
04	296	4: An-Nisaa, 5: Al-Maaida
05	165	6: Al-An'aam
06	206	7: Al-A'raaf
07	204	8: Al-Anfaal, 9: At-Tawba
08	232	10: Yunus, 11: Hud
09	154	12: Yusuf, 13: Ar-Ra'd
10	151	14: Ibrahim, 15: Al-Hijr
11	239	16: An-Nahl, 17: Al-Israa
12	208	18: Al-Kahf, 19: Maryam
13	247	20: Taa-Haa, 21: Al-Anbiyaa
14	196	22: Al-Hajj, 23: Al-Muminoon
15	141	24: An-Noor, 25: Al-Furqaan
16	227	26: Ash-Shu'araa

Chunk	Ayahs	Surahs
17	181	27: An-Naml, 28: Al-Qasas
18	163	29: Al-Ankaboot, 30: Ar-Room, 31: Luqman
19	157	32: As-Sajda, 33: Al-Ahzaab, 34: Saba
20	128	35: Faatir, 36: Yaseen
21	182	37: As-Saaffaat
22	163	38: Saad, 39: Az-Zumar
23	139	40: Ghafir, 41: Fussilat
24	142	42: Ash-Shura, 43: Az-Zukhruf
25	169	44: Ad-Dukhaan, 45: Al-Jaathiya, 46: Al-Ahqaf, 47: Muhammad
26	152	48: Al-Fath, 49: Al-Hujuraat, 50: Qaaf, 51: Adh-Dhaariyat
27	166	52: At-Tur, 53: An-Najm, 54: Al-Qamar
28	174	55: Ar-Rahmaan, 56: Al-Waaqia
29	166	57-66 (10 short Medinan surahs)
30	178	67: Al-Mulk, 68: Al-Qalam, 69: Al-Haaqqa, 70: Al-Ma'aarij
31	172	71: Nooh, 72: Al-Jinn, 73: Al-Muzzammil, 74: Al-Muddaththir, 75: Al-Qiyaama
32	167	76: Al-Insaan, 77: Al-Mursalaat, 78: An-Naba, 79: An-Naazi'aat
33	173	80-85 (Abasa through Al-Burooj)
34	175	86-95 (At-Taariq through At-Tin)
35	130	96-114 (Al-Alaq through An-Naas)

Extraction Targets Per Chunk

Each chunk extracts 6 structured sections:

1. **Theological themes and relationships** — 3-7 major themes, with verse citations [surah:ayah] and cross-surah connections
2. **Linguistic patterns** — repetition (verbatim counts), parallelism, chiasm/ring structures, refrains, oath structures

3. **Numerical observations** — verse counts, 19-divisibility checks, surah+verse sums, notable word frequency counts
4. **Cross-references and intertextuality** — verse echoes, shared prophet narratives, bookend relationships
5. **Distinctive vocabulary and phrases** — 5-10 unique terms, hapax legomena candidates, technical theological terms
6. **Chunk summary** — 3-5 sentence spiritual arc of the chunk

---

Test Chunk Results

Chunk 1 — Al-Faatiha (7 ayahs)

- **Time:** 180.7s total (151.1s generation)
- **Output:** 1,398 tokens, 5,484 chars
- **Speed:** 9.3 tok/s
- **Quality:** Excellent. Correctly identified chiasm A-B-C structure, parallelism in "Thee alone we worship / Thee alone we ask", themes of tawhid, divine attributes, worship, guidance. Cited verse numbers correctly.
- **Prompt tokens:** 865

Chunk 2 — Al-Baqara (286 ayahs)

- **Time:** 222.0s total (167.0s generation)
- **Output:** 1,441 tokens, 5,340 chars
- **Speed:** 8.6 tok/s
- **Quality:** Good. Identified monotheism, divine justice, faith+actions, divine guidance, eschatology. Detected refrain "O our Sustainer!" and parallelism in 2:277. **Issue noted:** Prompt was ~20K tokens but num_ctx was 8192 → model saw only the last portion of Al-Baqara. Fixed in script update (dynamic num_ctx up to 32768).
- **Prompt tokens:** 8,192 (was capped at context limit — now fixed)

Chunk 3 — Aal-i-Imraan (200 ayahs)

- **Time:** 191.5s total (136.6s generation)
- **Output:** 1,284 tokens, 4,915 chars
- **Speed:** 9.4 tok/s
- **Quality:** Good. Identified 5 themes (monotheism, prophethood, accountability, patience, hypocrisy), refrain "And God is aware of all that you do" at [3:154] and [3:160], cross-reference to Al-Baqara 2:208 and An-Nisa 4:173.
- **Prompt tokens:** 8,192 (same issue — fixed for future runs)

Fix Applied After Test: Dynamic Context Window

Problem: `num_ctx: 8192` was hardcoded. Large surahs (Al-Baqara = ~20K prompt tokens) had their text truncated.

Fix in `ollama-chunk-runner.js`:

```
const promptTokenEst = Math.ceil(prompt.length / 3.5);
const outputBudget   = 4096;
const numCtx = Math.min(32768, Math.max(8192, promptTokenEst + outputBudget + 512));
```

This dynamically sizes the context window to fit the full prompt + output budget, capped at 32768 (qwen2.5-coder:32b max). The M3 Ultra has sufficient RAM for 32K context on a 19GB model.

Implication: Large chunks (Al-Baqara, An-Nisaa+Al-Maaida, etc.) should now receive their full text. Chunks 2-4 should be re-run after clearing the existing output files if full-text analysis is required.

Time Estimates

Based on 3 test runs with qwen2.5-coder:32b:

Metric	Value
Chunk 1 (7 ayahs)	181s
Chunk 2 (286 ayahs)	222s
Chunk 3 (200 ayahs)	192s
Average per chunk	~198s (~3.3 min)
35 chunks × 198s	~115 minutes (~1.9 hours)

Revised estimate with dynamic context fix:

- Very large chunks (>200 ayahs) will take longer due to increased context loading
- Estimated 240-300s for the largest chunks (02, 04, 06, 07, 08, etc.)
- Conservative full-pass estimate: **2.5 - 3 hours**

To run the full pass (resumes from chunk 4 onward, chunks 1-3 already done):

```
bash ~/system/context/quran/ollama-full-pass.sh 4 35
```

To run from the beginning (chunks 2+3 will be skipped due to resume logic):

```
bash ~/system/context/quran/ollama-full-pass.sh
```

To re-run chunks 2-3 with the context fix (delete existing files first):

```
rm ~/system/context/quran/ollama-analysis/chunk-02.md
rm ~/system/context/quran/ollama-analysis/chunk-03.md
bash ~/system/context/quran/ollama-full-pass.sh 2 3
```

Files Created

File	Purpose
`~/system/context/quran/ollama-full-pass.sh`	Main orchestrator shell script
`~/system/context/quran/ollama-chunk-runner.js`	Node.js Ollama caller + output formatter
`~/system/context/quran/ollama-analysis/chunk-01.md`	Al-Faatiha analysis
`~/system/context/quran/ollama-analysis/chunk-02.md`	Al-Baqara analysis (partial — context limit)
`~/system/context/quran/ollama-analysis/chunk-03.md`	Aal-i-Imraan analysis (partial — context limit)
`~/system/context/quran/ollama-analysis/manifest.json`	Machine-readable progress tracker
`~/system/context/quran/ollama-analysis/run.log`	Run log for resume diagnostics

Issues Encountered

Issue 1: Context window truncation on large surahs (FIXED)

- **Problem:** `num_ctx: 8192` caused Al-Baqara's 20K-token prompt to be truncated to last ~8K tokens only. Model analyzed tail end of the surah rather than full text.
- **Fix:** Dynamic `num_ctx` calculation in `ollama-chunk-runner.js` — scales up to 32768.
- **Recommendation:** Re-run chunks 2 and 3 for full-text analysis.

Issue 2: Model hallucinated "286 divisible by 19" (minor)

- **Problem:** In chunk 2 analysis, model stated "The number of verses (286) is divisible by 19, a significant number in Quranic research." — 286 / 19 = 15.05, not divisible.
- **Diagnosis:** Model pattern-matched "286 and 19" association without checking arithmetic. Classic hallucination pattern.
- **Mitigation:** The prompt explicitly says to "flag any 19-related patterns" — but does not ask the model to verify arithmetic. A post-processing verification step could check all divisibility claims.
- **Recommendation:** Add a numerical fact-check pass as a separate script that verifies divisibility claims.

Issue 3: Chunk 1 is too small (7 ayahs)

- **Observation:** Al-Faatiha (7 ayahs) produced good analysis but is a single very short surah. The 181s processing time is dominated by model loading (~30s) + context construction, not actual content.
- **No fix needed** — Al-Faatiha is always analyzed alone due to its unique status as the opening prayer.

Quality Assessment

Section	Chunk 1	Chunk 2	Chunk 3
Theological themes	Excellent	Good	Good
Linguistic patterns	Good	Good	Good
Numerical observations	Adequate	Poor (hallucination)	Adequate
Cross-references	Good	Adequate	Good
Distinctive vocab	Good	Adequate	Adequate
Chunk summary	Excellent	Good	Good
Overall	**A**	**B-**	**B+**

The quality degrades slightly for large surahs due to context truncation. After the fix, chunks 2+ should reach A/B+ quality consistently.

Next Steps

1. Re-run chunks 2-3 after deleting existing files (context fix)
2. Run full pass chunks 4-35: `bash ~/system/context/quran/ollama-full-pass.sh 4 35`
3. After completion: write synthesis script that aggregates cross-chunk patterns
4. Optional: second pass with llama3.1:8b for comparison on selected chunks
5. Index all 35 chunk outputs in BookStack under Knowledge Base → Quran Research

---

Revision #3
Created 2026-03-05 05:18:47 UTC by John
Updated 2026-05-31 20:05:00 UTC by John