Comparative Analysis of Model Outputs Under Different Parameters

Test ID	max_tokens	temperature	Summary Label
T1	300	0.0	⚙️ Ultra-deterministic, concise
T2	300	0.1	🧊 Slight variation, still concise
T3	500	0.2	🔍 Balanced, more content depth
T4	700	0.2	📘 Extended, consistent detail
T5	700	0.3	🌿 Extended, with slight expressivity

1. Answer Completeness & Accuracy

Test	Sinhala QA	Constitution-based Qs	International Qs	Complex Qs (e.g., President's powers)
T1	✅ Brief	⚠️ Truncated answers	✅ Correct	❌ Often cut off
T2	✅ Brief	⚠️ Slightly better	✅ Correct	⚠️ Still lacks elaboration
T3	✅ Full	✅ Complete	✅ Correct	✅ Sufficient legal nuance
T4	✅ Detailed	✅ Extended citations	✅ Correct	✅ Broader coverage
T5	✅ Detailed	✅ Same as T4	✅ Correct	✅ Slightly more natural tone

- From T3 upward, answers became significantly more **informative** and **legally complete**.
- T4 and T5 handled questions like **Presidential pardon powers** with **caution** and **legal qualifiers**, reducing hallucination risk.

2. Language Fluency & Naturalness (Sinhala)

Test	Formal Sinhala	Expressiveness	Clarity
T1	✅ Very formal	❌ Flat	✅ Clear
T2	✅ Formal	⚠️ Slight nuance	✅ Clear
T3	✅ Formal	⚠️ Modestly rich	✅ Clear
T4	✅ Formal	⚠️ Stable	✅ Clear
T5	✅ Formal	✅ Slightly richer	✅ Natural

- T5 introduced **mild stylistic elegance** in Sinhala (e.g., contextual judgment).
- T1–T2 are best for rigid legal tone; T5 for educational or public-facing interfaces.

3. Consistency Across Similar Questions

Test	Consistency in phrasing	Logical structure	Contradictions
T1	✅ High	✅ Strong	❌ Some omissions
T2	✅ High	✅ Strong	❌ Same as T1
T3	✅ High	✅ Improved	✅ Handled nuance
T4	✅ High	✅ Detailed	✅ Accurate
T5	✅ High	✅ Slightly more human	✅ Same

- From T3 upward, answers gain **logical scaffolding** (e.g., bulleting, qualification).
- T5 avoids contradictions while being more engaging.

4 Structural Presentation

Test	Use of Headings / Bullets	Segmentation	Readability
T1	❌ None	❌ One block	⚠️ Dense
T2	⚠️ Rarely	⚠️ Limited	⚠️ Average
T3	✅ Some bullets	✅ Structured	✅ Improved
T4	✅ Consistent structure	✅ Clear	✅ High
T5	✅ Same as T4	✅ Clear	✅ High + natural

Overall Ratings Table

Criteria	T1	T2	T3	T4	T5
Accuracy	✅	✅	✅	✅	✅
Completeness	❌	⚠️	✅	✅	✅
Legal Sensitivity	⚠️	⚠️	✅	✅	✅
Retrieval Use (RAG)	❌	❌	⚠️	⚠️	⚠️
Language Fluency	✅	✅	✅	✅	✅+
Natural Tone	❌	⚠️	⚠️	⚠️	✅
Stylistic Variation	❌	⚠️	⚠️	✅	✅+
Best Use Case	Static bots	Legal chatbots	Civic Q&A	Gov platforms	Public legal education

RAG Quality Evaluation Table

Test ID	Retrieval Presence	Relevance of Retrieved Chunks	Integration into Answer	Hallucination Reduction	Observed Issues
T1	✅ Yes	❌ Very low	❌ Not used	❌ No impact	Context mostly ignored or irrelevant (e.g., metadata chunks, wrong sections)
T2	✅ Yes	❌ Poor	❌ Not used	❌ No impact	Slight variation from T1, but still poor chunk-to-query alignment
T3	✅ Yes	⚠️ Mixed (some partial hits)	⚠️ Weak references	⚠️ Slight reduction	Answer depends on model memory; retrieved content included but not relied upon
T4	✅ Yes	⚠️ Slightly better than T3	⚠️ Referenced indirectly	⚠️ Moderate impact	Mentions clauses (e.g., 43–46), but full arguments built by model internally
T5	✅ Yes	⚠️ Same as T4	✅ Blended better	✅ Slight improvement	Slightly more organic use of retrieved sentences, but still requires tighter matching

RAG Dimensions Breakdown

Metric	T1	T2	T3	T4	T5
Chunk Relevance	1/5	2/5	3/5	3.5/5	3.5/5
Answer Dependency on RAG	❌	❌	⚠️	⚠️	✅

Use of Clause/Article IDs	❌	❌	✅	✅	✅
Overlap with Correct Answer	❌	❌	⚠️	✅	✅
Degree of Hallucination	High	High	Medium	Low	Low

- ✅ **T5** is the only test where RAG context **blends naturally** into the answer and reinforces confidence in legal accuracy.
- ⚠️ **T3 & T4** refer to legal articles and display improved structure, but they do not *depend* heavily on the retrieval results — likely using model-internal knowledge.
- ❌ **T1 & T2** retrieval is essentially **wasted**, likely due to low chunk relevance and short max tokens.