

# MSIS 549 HW2: Benchmark Appendix

## 1. Evaluation Methodology

**Methods Used:** Human Rubric Scoring (Method 1) + Baseline Comparison (Method 3)

**Baseline:** Single-prompt request to GPT-4: "Write 6 LinkedIn posts about SQL Query Performance for Data Engineers and Business Stakeholders."

**Agentic System:** Full 9-skill pipeline (Intent Discovery → Content Strategist → Draft Architect → Voice Refiner → Engagement Optimizer → Quality Review).

**Scoring:** All posts scored by the author (human rubric) on a 1-5 scale. Prompts and settings were frozen across all test cases.

## 2. Scoring Rubric (Frozen)

Score	Anchor Description
5	Excellent — High-impact, strategically deep, sounds like a real senior leader. Specific, actionable, and well-formatted for LinkedIn.
4	Good — Professional and insightful, minor gaps in specificity or voice consistency.
3	Satisfactory — Clear and competent, but reads as generic thought leadership. Lacks personal anecdotes or unique strategic depth.
2	Below Average — Surface-level, uses AI cliches, could be about any topic. Weak hook or CTA.
1	Poor — Obviously AI-generated, no strategic depth, disconnected from audience needs.

## 3. Metrics (5 Dimensions)

#	Metric	What It Measures
1	Actionability	Does the post give the reader something concrete to do?
2	Voice Consistency	Does this sound like the same leader across all 6 posts?
3	Strategic Depth	Does it demonstrate genuine expertise and insider knowledge?
4	Narrative Cohesion	Does it build on previous weeks and tease the next?
5	LinkedIn Optimization	Strong hook, clean formatting, hashtags, visual suggestion?

## 4. Test Case 1: SQL Query Performance (Primary)

**Input (Strategic Intent):**

- Topic:** Optimizing SQL Query Performance
- Audience:** Data Engineers & Business Stakeholders
- Core Message:** Continuous improvement in SQL is essential for AI-readiness

- **Anecdote:** Business team frustrated when data wasn't in sync with AI models; "near real-time" became a requirement
- **CTA:** Rethink your data platform strategy; follow for ongoing insights
- **Tone:** Provocative + Educational

#### 4a. Baseline Output (Single-Prompt GPT-4)

Prompt: "Write 6 LinkedIn posts about SQL Query Performance for Data Engineers and Business Stakeholders."

##### Baseline Post 1 (excerpt):

"In today's data-driven world, SQL query performance is more important than ever. Here are 5 tips to optimize your queries: 1) Use indexes wisely 2) Avoid SELECT \* 3) Optimize JOINS 4) Use query execution plans 5) Consider partitioning. What are your favorite SQL optimization tips? Drop them in the comments! #SQL #DataEngineering"

##### Baseline Post 3 (excerpt):

"Let's dive into a common challenge: slow queries in production. Many teams struggle with this issue. Here are some best practices to address it: First, analyze your execution plan. Second, check for missing indexes. Third, consider caching strategies. What other approaches have worked for your team? #Database #Performance"

**Baseline Assessment:** All 6 posts follow the same "here are X tips" pattern. No narrative arc, no personal anecdotes, no strategic depth. Generic hooks ("In today's..."), generic CTAs ("Drop in the comments"). Posts could be reordered without any loss of meaning.

#### 4b. Agentic Output (This System)

##### Agentic Post 1 — "The Good Enough Trap" (excerpt):

"Is your data platform actually 'good enough,' or is it just holding you back? Traditional SQL performance is no longer a technical detail — it's a strategic bottleneck. Many teams settle for 'good enough' query speeds, but as business demands shift toward real-time insights, that complacency becomes a liability... Performance is the foundation of agility."

##### Agentic Post 3 — "Near Real-Time is the New Baseline" (excerpt):

"The day 'near real-time' became a requirement, not a request. I remember a time when business teams were happy with daily reports. Those days are gone. Recently, I saw a business team's frustration when their data wasn't in sync with the AI models they were using for decision-making... Stakeholder expectations are driven by the fastest tool in their kit."

#### 4c. Scoring — Test Case 1

##### Baseline (Single-Prompt GPT-4):

Post	Actionability	Voice	Depth	Cohesion	LinkedIn	Avg
Post 1	3	2	2	1	3	2.2
Post 2	3	2	2	1	3	2.2
Post 3	2	3	2	1	2	2.0
Post 4	3	2	2	1	3	2.2
Post 5	2	2	2	1	2	1.8

Post 6	2	2	1	1	2	1.6
Avg	2.5	2.2	1.8	1.0	2.5	2.0

Agentic System (This System):

Post	Actionability	Voice	Depth	Cohesion	LinkedIn	Avg
Week 1: Good Enough Trap	4	5	4	5	5	4.6
Week 2: Instant Satisfaction Gap	4	5	4	5	4	4.4
Week 3: Near Real-Time	5	5	5	5	4	4.8
Week 4: Technical Levers	5	4	5	4	5	4.6
Week 5: AI-Ready Infrastructure	3	4	4	5	4	4.0
Week 6: Strategic Pivot	4	5	4	5	4	4.4
Avg	4.2	4.7	4.3	4.8	4.3	4.5

5. Test Case 2: Enhancing Data Cleanliness

Input (Strategic Intent):

- **Topic:** Enhancing Data Cleanliness
- **Audience:** Data Analysts & Product Managers
- **Core Message:** Data quality is the foundation of trustworthy AI — clean data isn't optional, it's strategic
- **Anecdote:** A product launch delayed by 2 weeks because the ML model was trained on dirty customer data
- **CTA:** Audit your top data sources for quality this quarter
- **Tone:** Educational + Empathetic

5a. Agentic Output Summary (Test Case 2)

The system generated a 6-week roadmap:

Week	Title
1	"The Dirty Data Tax" — cost of poor data quality
2	"The 80/20 Rule of Data Cleaning" — focus on highest-impact sources
3	"The Launch That Almost Wasn't" — personal anecdote
4	"5 Data Quality Checks Every Pipeline Needs" — tactical advice
5	"AI Can't Fix What You Won't Measure" — data quality for ML
6	"Building a Data Quality Culture" — organizational change CTA

5b. Scoring — Test Case 2

Post	Actionability	Voice	Depth	Cohesion	LinkedIn	Avg
------	---------------	-------	-------	----------	----------	-----

Week 1	4	4	4	5	4	4.2
Week 2	5	4	4	4	4	4.2
Week 3	4	5	5	5	4	4.6
Week 4	5	4	4	4	5	4.4
Week 5	3	4	4	4	3	3.6
Week 6	4	4	3	5	4	4.0
<b>Avg</b>	<b>4.2</b>	<b>4.2</b>	<b>4.0</b>	<b>4.5</b>	<b>4.0</b>	<b>4.2</b>

## 6. Edge Case: Highly Technical Topic

**Input:** "Implementing Write-Ahead Logging in Distributed Database Consensus Protocols"

- **Audience:** Database kernel engineers
- **Tone:** Data-driven

**Result:** The system handled this well through Skills 1-2 (the interview forced clarification of the audience and simplified the scope). However, Skill 4 (Voice Refiner) struggled — it over-simplified technical terminology that the audience would expect (e.g., replacing "WAL" with "write-ahead logging" every time, which felt condescending to kernel engineers).

**Score:** 3.4/5.0 average — acceptable but weaker on Voice Consistency (2.5) because the "senior leader" voice template doesn't fit highly specialized technical audiences well.

**Lesson:** The skill set is optimized for Manager/Director-level audiences. For deeply technical niche audiences, Skill 4 needs a "technical expert" mode that preserves jargon rather than simplifying it.

## 7. Ambiguous Case: Vague Business Goal

**Input:** "I want to write about making things better with data"

- No specific audience, message, anecdote, or tone provided initially.

**Result:** Skill 1 (Intent Discovery) handled this well — the 5-question interview forced the user to clarify:

- Audience: "Mid-level managers in retail"
- Core message: "Data-driven decision making reduces operational waste"
- Anecdote: Prompted user to share one; user provided a story about inventory forecasting
- Tone: Educational

The downstream skills produced a reasonable series, scoring 3.8/5.0 average. The weakest area was Strategic Depth (3.2) — because the initial topic was so broad, the series stayed at a high level rather than diving deep.

**Lesson:** The interview (Skill 1) is the most critical skill. Vague inputs can be rescued but result in shallower content. A possible improvement would be adding a "topic sharpening" step between Skills 1 and 2.

## 8. Aggregate Results

Test Case	Actionability	Voice	Depth	Cohesion	LinkedIn	<b>Overall</b>
-----------	---------------	-------	-------	----------	----------	----------------

TC1: SQL Performance	4.2	4.7	4.3	4.8	4.3	4.5
TC2: Data Cleanliness	4.2	4.2	4.0	4.5	4.0	4.2
Edge: WAL Protocol	3.5	2.5	4.0	3.8	3.2	3.4
Ambiguous: Vague Input	3.5	3.8	3.2	4.0	3.5	3.6
Baseline: Single-Prompt	2.5	2.2	1.8	1.0	2.5	2.0

Key Findings

- **Agentic system outperformed baseline by +2.1 points** on the primary test case (4.5 vs 2.0).
- **Biggest improvement:** Narrative Cohesion (+3.8 over baseline) — the 6-week roadmap ensures posts build on each other, which a single prompt cannot achieve.
- **Strongest metric:** Narrative Cohesion (avg 4.3 across all agentic runs) — the structured arc from Skill 2 is consistently effective.
- **Weakest metric:** Voice Consistency on edge cases (2.5) — the voice refiner needs audience-specific modes.

Worst Failure

**Edge Case, Week 5:** The post on "consensus protocol implications for AI workloads" scored 2.0 on Voice Consistency. The Skill 4 refiner replaced precise technical terminology with strategic business language, which felt patronizing to the target audience of database kernel engineers. The anecdote about "leadership decisions" felt forced in a deeply technical series.

**Root Cause:** Skill 4 is hard-coded to a "Manager/Director" voice. It needs a configuration parameter to select between voice profiles (e.g., `executive` , `technical-expert` , `practitioner` ).

9. Reproducibility Notes

- **LLM:** Manus AI (Claude-based) for primary runs; GPT-4 for baseline
- **Prompts:** Frozen as documented in each skill's `.md` file
- **Settings:** Default temperature, no custom parameters
- **Evaluator:** Human scoring by the author (single rater)
- **Limitation:** Single rater introduces subjective bias. An improvement would be inter-rater reliability with 2+ evaluators.