WWTP Engineering Benchmark

Wastewater Treatment Plant LLM Evaluation Suite

10 Tasks | Domain Expert Knowledge | Real-World Scenarios

Author: Mehmet ISIK

Kaggle Grandmaster | WWTP Operations Expert

December 2025

Executive Summary

This benchmark suite evaluates Large Language Models on their ability to handle real-world wastewater treatment plant engineering challenges. Unlike generic benchmarks, these tasks require domain-specific knowledge that is typically not found in standard training corpora.

The benchmark consists of 10 tasks covering material selection, root cause analysis, process chain thinking, safety protocols, biological processes, laboratory procedures, and industrial electrical standards.

Key Finding: Tasks requiring real field experience (such as digester walkway material selection and biogas desulfurization recovery) showed significantly lower pass rates (1/6), demonstrating that these scenarios effectively test knowledge beyond standard training data.

Consolidated Results

Performance of 6 models across all 10 tasks:

Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Total
Claude Opus 4.5	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	**8/10**
Claude Sonnet 4.5	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓	**8/10**
Gemini 2.5 Flash	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓	**8/10**
Gemini 3 Pro	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	**7/10**
Qwen 3 Coder 480B	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗	**5/10**
Gemini 2.0 Flash	✓	✗	✗	✓	✓	✗	✓	✗	✓	✓	**6/10**
Pass Rate	**6/6**	**5/6**	**5/6**	**1/6**	**5/6**	**1/6**	**6/6**	**5/6**	**4/6**	**4/6**	

Task Legend: T1=Equipment Material | T2=Root Cause | T3=Process Chain | T4=Walkway Material | T5=Confined Space | T6=Biogas Recovery | T7=Sample Preservation | T8=Tool Selection | T9=Emergency Flocculation | T10=SCADA Cabling

Task Summaries

Task 1: Equipment Material Selection

Scenario: Select corrosion-resistant material for submersible pump in H2S environment.

Correct Answer: Stainless Steel 316 (not Cast Iron)

Result: **6/6 PASS** - Baseline task

Task 2: Root Cause Analysis

Scenario: Diagnose pump failure from symptoms: overheating, reduced flow, high current.

Correct Answer: Clogged impeller

Result: **5/6 PASS** - Gemini 2.0 Flash failed

Task 3: Process Chain Analysis

Scenario: Identify root cause of decanter vibration and scratched scroll blades.

Correct Answer: Grit removal system failure (upstream thinking)

Result: **5/6 PASS**

Task 4: Digester Walkway Material

Scenario: Select material for walkway gratings at 15m height in H2S/CH4 environment.

Correct Answer: Stainless Steel 316 (not FRP)

The Trap: FRP is corrosion-resistant but BRITTLE - unsafe for personnel walkways at height.

Result: **1/6 PASS** - HARDEST TASK! Only Gemini 2.0 Flash passed.

Task 5: Confined Space Emergency

Scenario: Select correct safety protocol for confined space entry.

Result: **5/6 PASS**

Task 6: Biogas Desulfurization Recovery

Scenario: Identify bacterial source for re-inoculating inhibited biological sulfur removal unit.

Correct Answer: Anaerobic digester sludge (not activated sludge)

The Trap: Activated sludge has diverse bacteria but AEROBIC. Sulfur bacteria live in anaerobic digesters where H2S is produced.

Result: **1/6 PASS** - Only Claude Opus 4.5 passed!

Task 7: Sample Preservation Protocol

Scenario: Select proper storage conditions for wastewater sample over weekend.

Correct Answer: +4C refrigeration, closed container, dark conditions

Result: **6/6 PASS** - Baseline task

Task 8: Confined Space Tool Selection

Scenario: Select wrench type for work in sludge manhole with explosion risk.

Correct Answer: Brass wrench (non-sparking)

Result: **5/6 PASS**

Task 9: Emergency Flocculation

Scenario: Select emergency intervention for activated sludge settling failure.

Correct Answer: Polyelectrolyte (polymer) dosing

The Trap: FeCl3 is chemical coagulant that harms bacteria. Polymer is physical flocculant - safe for biology.

Result: **4/6 PASS**

Task 10: SCADA Cabling Standards

Scenario: Select cabling approach for SCADA integration over long distance.

Correct Answer: Separate power and signal cables in SEPARATE conduits

The Trap: Same conduit appears cost-effective but causes EMI interference - violates IEC standards.

Result: **4/6 PASS** - Claude Opus and Qwen prioritized cost over reliability

Key Findings

1. Safety vs Technical Optimization

In Task 4, 5/6 models chose FRP for corrosion resistance, ignoring that FRP is brittle and dangerous for personnel walkways at 15m height. When human safety is involved, material behavior under failure matters more than chemical resistance.

2. Aerobic vs Anaerobic Process Knowledge

Task 6 revealed a knowledge gap: only 1/6 models understood that sulfur-oxidizing bacteria exist in anaerobic digesters. The key insight - 'where H2S is produced, sulfur bacteria exist' - comes from field experience, not textbooks.

3. Cost vs Reliability Trade-offs

In Task 10, even Claude Opus 4.5 chose cost optimization over industrial EMI standards. Models may not adequately weight long-term reliability over short-term savings in industrial contexts.

4. Process Chain Thinking

Task 3 required upstream thinking: decanter problems caused by grit removal failure. Models that focus only on the immediate symptom miss root causes that require understanding the entire process flow.

Conclusion

This benchmark demonstrates that real-world industrial engineering expertise remains challenging for LLMs. Tasks requiring field experience, safety-critical judgment, and process-level understanding showed significantly lower pass rates than standard technical questions.

The most valuable benchmarks test knowledge unlikely to be present in training data - specifically, operational insights from years of hands-on experience in specialized domains.

Author: Mehmet ISIK - Kaggle Grandmaster, WWTP Operations Expert