WB GPT Testing Report

Pressure Testing & Review Findings

Prepared by:	Gowtham Kumar Kanchi
Date:	February 3, 2026
Version:	1.0
Project:	WB GPT - Winning Business Allocation Calculator

Executive Summary

Following our alignment meeting on January 30, 2026, I conducted comprehensive pressure testing on the WB GPT to evaluate its performance against the core requirements discussed. This report documents the testing methodology, results, and recommendations for improvement.

Overall Assessment

Category	Status
Gating System	✓ **Working Well**
Ambiguous Input Handling	✓ **Working Well**
Individual Cap Logic (100%)	✓ **Working Well**
Team Scaling Math	✓ **Working Well**
JSON Output Format	✓ **Working Well**
Repeatability/Determinism	✗ **CRITICAL ISSUE**
Grid Level Mapping Accuracy	⚠ **Needs Improvement**

Test Results Summary

Test	Purpose	Expected	Actual	Result
Test 1	Repeatability	Identical to Minh's output	Different	**FAIL**
Test 2A	Team scaling (275%→200%)	Scale factor 0.727	0.727	**PASS**
Test 2B	Individual cap (150%→100%)	Dave capped, surplus tracked	Correct	**PASS**
Test 2C	Ambiguous inputs	GPT asks clarification	Correct	**PASS**
Test 3A	WB Guide Scenario 1	A=100%, B=50%, C=50%	Over-credited	**FAIL**
Test 3B	WB Guide Scenario 6	A=50%, B=75%, C=75%	Close	**PARTIAL**
Test 3C	Execution vs Pitch	B gets 0-25% max	B got 0%	**PASS**

Critical Issue #1: Repeatability Failure

This is the most critical issue identified during testing.

When running the exact same inputs that Minh used in her original test, the GPT produced different percentage allocations. This defeats the core requirement of determinism - same inputs must produce same outputs every time.

Comparison: Minh's Test vs My Test (Identical Inputs)

Person	Minh's Test (Base WB%)	My Test (Base WB%)	Difference
Minh	50% (In-take)	75% (Proactive)	+25%
Gowtham	25% (MSA)	50% (MSA)	+25%
Dany	25% (Participant)	50% (Helpful)	+25%
Bryan	100% (Open/Close)	100% (Open/Close)	0%
Total Base	**200%**	**275%**	**+75%**

Critical Issue #2: Over-Splitting Contributions

In Test 3A (WB Guide Scenario 1), the GPT split combined contribution descriptions into multiple separate entries, significantly inflating the WB percentages.

Example: Person A's Contribution

Input description: "has history with client and expertise of industry - foundational prior work and mastery expertise"

Expected (per WB Guide): 100% (single combined contribution)

Actual GPT mapping: 225% (split into 3 separate entries at 75% each)

- Existing client or relationships → 75%
- Prior work related → 75%
- Helping with a pitch, providing expertise → 75%

Impact: This over-splitting caused the total team allocation to balloon from the expected 200% to 450%, requiring aggressive scaling that distorts the relative contributions.

What's Working Well

- **Gating System:** All 5 gates work correctly with VERIFIED checkpoints
- **Ambiguous Input Detection:** Correctly refuses vague contributions and asks for specifics
- **Individual Cap Enforcement:** Correctly caps at 100% and tracks house/surplus
- **Team Scaling Math:** Scale factor calculations are accurate
- **Execution vs Pitch Detection:** Correctly identifies and rejects non-contributing roles
- **JSON Output Format:** Valid, complete, matches required schema

Recommendations

Priority 1: Fix Repeatability (Critical)

1. Add explicit mapping examples in instructions showing exact input → output patterns
2. When contribution descriptions combine multiple aspects, instruct the GPT to map to the single highest applicable level, not multiple entries
3. Add a decision tree or flowchart for grid level selection

Priority 2: Add Common Scenarios as Reference

The WB Calculations PDF contains a Common Scenarios table that could serve as explicit lookup. Adding these as reference cases could improve consistency:

Scenario	Pattern	Expected Allocation
1	Existing client + pitch expertise (3 people)	A=100%, B=50%, C=50%
6	Existing relationship + mastery expertise (3 people)	A=50%, B=75%, C=75%
12	Reputation + mastery expertise (2 people)	A=100%, B=100%

Priority 3: Clarify Grid Level Boundaries

Add clearer decision criteria for distinguishing between similar levels, particularly:

- 100% (Mastery) vs 75% (Meaningful) for expertise contributions
- 75% (Proactive) vs 50% (In-take) for relationship contributions
- 100% (Reputational Leadership) vs 50% (Proxy) for reputation contributions

Test Conversation Links

All test conversations are available for review at the links provided in the accompanying email.