WB GPT Testing Report

Pressure Testing & Review Findings

Prepared by: Gowtham 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**

Testing Methodology

Test Categories

1. **Repeatability Test** - Same inputs as Minh's original test to verify deterministic behavior
2. **Edge Case Tests** - Team scaling, individual caps, ambiguous inputs
3. **WB Guide Scenario Tests** - Scenarios from WB User Guide pages 12-13 to verify mapping accuracy

Test Environment

- GPT: WB GPT (V2 Compact Instructions)
- Knowledge File: WB User's Guide 2024
- Mode: All tests conducted using Mode A (single text block)

Test Results Detail

Test 1: Repeatability Test

Objective: Verify same inputs produce same outputs (determinism)

Input Used:

```
client: acme, minh: introduced client, gowtham: created contracts,
dany: setup meetings and candidates, bryan: closed contract, got sign off
WB_$for200 = 90000, ScalePct = 80
```

Comparison: Minh's Test vs My Test

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%**

Result: ❌ FAIL

Impact: This is a critical failure. The core requirement from our meeting was determinism - same inputs must produce same outputs every time. This test proves the GPT is not currently deterministic.

Test 2A: Team Scaling Test (Over 200%)

Objective: Verify team scaling when total exceeds 200%

Input Used:

```
client: TechCorp
Alice: acquired new client through proactive BD and market activity that led to pitch and win
Bob: prior work was foundational - excellence of prior work was cited as reason for new work
Charlie: expertise was critical and determinative - RRA won hands down because of his mastery expertise
```

Expected vs Actual:

Check	Expected	Actual	Status
Alice Base WB%	100%	100%	✅
Bob Base WB%	75%	75%	✅
Charlie Base WB%	100%	100%	✅
Total Base	275%	275%	✅
Scale Factor	0.7273	0.7273	✅
Alice Scaled	72.73%	72.73%	✅
Bob Scaled	54.55%	54.55%	✅
Charlie Scaled	72.73%	72.73%	✅
Team Total After Scaling	200%	200%	✅

Result: ✅ **PASS**

Conclusion: Team scaling math works correctly.

Test 2B: Individual Cap Test (Over 100% Individual)

Objective: Verify individual cap enforcement and house/surplus tracking

Input Used:

```
client: MegaCorp
Dave: acquired new client through traditional BD AND was asked by name due to market reputation - reputational leadership
Eve: developed complex search strategy requiring customization - helpful expertise
```

Expected vs Actual:

Check	Expected	Actual	Status
Dave Base WB%	200% (100+100)	150% (100+50)	⚠️ See note
Dave Capped	100%	100%	✅
Eve Base/Capped	50%	50%	✅
House/Surplus	100%	50%	⚠️ See note
Team Total	150%	150%	✅
Scale Factor	1.0	1.0	✅

Result: ✅ **PASS** (with mapping note)

Note: Per WB Guide Page 10, "Reputational Leadership" (being asked by name due to market reputation) should be **100%**, not 50%. The GPT mapped it as 50% (Proxy level) instead of 100% (Reputational Leadership). The cap/scaling mechanics worked correctly, but the initial mapping was under-credited.

Test 2C: Ambiguous Input Test

Objective: Verify GPT challenges vague or unclear contributions

Input Used:

```
client: GlobalInc
Frank: he knows the client
Grace: she helped out
```

Expected Behavior: GPT should refuse to proceed and ask for clarification

Actual Behavior:

- ✅ GPT flagged "he knows the client" as **UNCLEAR**
- ✅ GPT flagged "she helped out" as **NO/UNCLEAR**
- ✅ GPT refused to assign WB% without clarification
- ✅ GPT explained what specific information is needed
- ✅ GPT blocked VERIFIED response until clarification provided

Result: ✅ **PASS**

GPT Response (excerpt):

> "I cannot assign WB% to either person unless you clarify what they actually did that directly influenced why GlobalInc awarded the work."

Conclusion: This is exactly the behavior we want - the GPT correctly enforces the "direct contribution" requirement.

Test 3A: WB Guide Scenario 1 (Page 12)

Objective: Verify GPT matches WB User Guide expected allocations

Scenario from WB Guide:

> Client that has done prior work with A asks us to pitch for a role in B's space. A has a history with Client and expertise of industry, B is strong in the capability/industry. C has done assessments in the space. We win based on A's history and expertise, B's extensive capability and C's insights. The expertise of B and C is equally balanced in the win.

Input Used:

```
client: ClientX, work type: Search, competitive: yes
A: has history with client and expertise of industry - foundational prior work and mastery expertise
B: strong in capability/industry - provided expertise that was equally balanced with C in securing the win
C: has done assessments in the space - provided expertise that was equally balanced with B in securing the win
```

Expected vs Actual (Base WB%):

Person	Expected (per Guide)	GPT Gave	Issue
A	100%	225% (3 × 75%)	Over-credited
B	50%	75%	Over-credited
C	50%	150% (2 × 75%)	Over-credited
Total	**200%**	**450%**	**+250%**

Result: ❌ **FAIL**

Root Cause Analysis:

The GPT is **splitting combined descriptions into multiple separate contributions**:

For Person A, the description "foundational prior work and mastery expertise" became:

1. Existing client or relationships → 75%
2. Prior work related → 75%
3. Helping with a pitch, providing expertise → 75%
4. **Total: 225%**

The WB Guide treats this as **one combined contribution** worth 100%, not three separate contributions totaling 225%.

Test 3B: WB Guide Scenario 6 (Page 13)

Objective: Verify GPT matches WB User Guide Scenario 6

Scenario from WB Guide:

> A receives a call from a prior client based on their history. The client is looking for work in an area that is not relevant to A who passes the opportunity onto consultants B and C. B and C pitch with A joining to provide support and knowledge of the client. It was the mastery expertise related to the opportunity that secured the win.

Expected Allocation (per Guide):

- A = 50% (Proactive + useful expertise)
- B = 75% (Mastery expertise)
- C = 75% (Mastery expertise)
- Total = 200%

Input Used:

```
client: ClientY
A: received call from prior client based on history but work is in area not relevant to A - passed opportunity to B and C - proactive introduction and joined pitch to provide support and client knowledge - useful expertise
B: pitched with mastery expertise that was critical to securing the win
C: pitched with mastery expertise that was critical to securing the win
```

Expected vs Actual:

Person	Expected (per Guide)	GPT Base WB%	Status
A	50%	50% (25%+25%)	✅ Correct
B	75%	100%	❌ Over-credited
C	75%	100%	❌ Over-credited
Total	**200%**	**250%**	Over by 50%

After Scaling (Scale Factor = 0.8):

Person	Expected Final	GPT Final	Variance
A	50%	40%	-10%
B	75%	80%	+5%
C	75%	80%	+5%

Result: ⚠️ PARTIAL PASS

Issue: The GPT mapped B and C's "mastery expertise that was critical" to 100% (Mastery level), but per Guide Scenario 6, they should be 75% each. The guide indicates the expertise was determinative but not sole - hence 75% not 100%.

Test 3C: Execution vs Pitch Contribution Test

Objective: Verify GPT correctly handles "Trading WB for execution" bad behavior

Scenario (from WB Guide Page 7 - Bad Behaviors):

> Consultant A and B won work with the client but needs an additional resource to execute. C is asked to execute, and the client agrees to have C execute after a brief meeting. C asks for equal WB with A and B.

Input Used:

```
client: TestClient
A: won the work and will execute
B: was asked to execute the project after we won - client agreed to have B execute after a brief meeting - B is asking for equal WB with A
```

Expected: B should get 0-25% max (per WB Guide), ideally 0%

Actual:

Check	Expected	Actual	Status
A (won the work)	100%	100%	✅
B (execution only)	0-25% max	0%	✅
GPT challenged B's request	Yes	Yes	✅
GPT explained reasoning	Yes	Yes	✅

Result: ✅ PASS

GPT Response (excerpt):

> "Under WB rules, execution after the work is already won does not automatically qualify for WB, unless the post-award involvement directly influenced: Keeping the work from falling apart, or A material change that was necessary for the client to proceed"

Conclusion: The GPT correctly identified and rejected the "Trading WB for execution" bad behavior pattern.

Summary of Issues

Critical Issues (Must Fix)

#	Issue	Description	Impact
1	**Non-deterministic outputs**	Same inputs produce different results on different runs	Defeats core purpose of the tool
2	**Over-splitting contributions**	Combined descriptions become multiple separate entries, inflating WB%	Causes unfair/incorrect allocations
3	**Inconsistent grid level selection**	Same contribution type maps to different percentages	Unpredictable results

Minor Issues

#	Issue	Description	Impact
4	Reputational Leadership mapping	Sometimes mapped to 50% instead of 100%	Under-credits reputation

#	Issue	Description	Impact
5	Mastery vs Meaningful confusion	Sometimes gives 100% when 75% is appropriate	Over-credits expertise

What's Working Well

Feature	Status	Notes
Gating System (Gates 0-5)	✅ Excellent	All gates work correctly with VERIFIED checkpoints
Ambiguous Input Detection	✅ Excellent	Correctly refuses vague contributions
Individual Cap Enforcement	✅ Working	Correctly caps at 100%
Team Scaling Math	✅ Working	Scale factor calculations are accurate
House/Surplus Tracking	✅ Working	Correctly tracks excess from caps
Execution vs Pitch Detection	✅ Excellent	Correctly identifies non-contributing roles
JSON Output Format	✅ Working	Valid, complete, matches required schema
Direct Influence Validation	✅ Working	Asks for confirmation on each contribution

Recommendations

Priority 1: Fix Repeatability (Critical)

Problem: The GPT produces different outputs for identical inputs.

Suggested Solutions:

1. Add explicit mapping examples in instructions showing exact input → output patterns
2. When contribution descriptions combine multiple aspects (e.g., "foundational prior work and mastery expertise"), 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
4. Consider adding temperature=0 or similar determinism settings if available in GPT configuration

Priority 2: Add Common Scenarios as Reference

The WB Calculations PDF contains a Common Scenarios table that could serve as explicit lookup:

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

Adding these as explicit reference cases could improve consistency.

Priority 3: Clarify Grid Level Boundaries

The GPT sometimes confuses similar levels. Add clearer decision criteria:

For Expertise:

- **100% (Mastery):** Expertise is the SOLE or PRIMARY reason for win - "RRA won hands down"
- **75% (Meaningful):** Expertise truly sets RRA apart but not sole factor
- **50% (Helpful):** Contributes meaningfully but others also contributed equally
- **25% (Useful):** Supporting role, not decisive

For Relationships:

- **100% (Pick-up):** Project not competitive due to relationship; OR creates new opportunities
- **75% (Proactive):** Actively introduces to new markets, creates opportunities
- **50% (In-take):** Brief intake and intro to colleagues who then pitch
- **0% (Pass on):** Simply passes lead to relevant colleagues

For Reputation:

- **100% (Reputational Leadership):** Asked for BY NAME due to reputation
- **50% (Proxy):** Not attending pitch but expertise demonstrated by others
- **25% (Creating awareness):** RRA included due to general reputation

Priority 4: Add Validation Against Guide Scenarios

Consider adding a validation step that compares calculated allocations against known WB Guide scenarios

before final output. If the allocation pattern matches a known scenario but percentages differ significantly, flag for review.

Report Prepared By: Gowtham Kanchi
Date: February 3, 2026