ANDREW TEMPLETON

Is your AI initiative a good DoG or bad DoG?

Definition Of Good

An org-deployable method and prompt to get consistent, high quality AI bets.

AI project proposal terms of engagement.

A CEO friend asked me one night:

Would a search relevancy change initiative actually help?

Two nerds. Ninety minutes. We got there - but it was hard. I needed a gut-check system every team could use, not just two people who happened to be nerds.

"Definition of Good" is half the battle.

```
Value  ≈  Backing  ×  Bite  ×  Bet
```

Any term at zero kills the case.

Backing	evidence ladder · how do you know?
Bite	non-obvious update · are you learning anything?
Bet	financial weight · who would bet money on it?

Standing on Popper, Shannon, Lindley, Raiffa.

Backing.

How do you know the claim? What would change your mind?

What it sounds like at each level:

"LLM tags are better."
Hunch. No falsifier; no change-my-mind observation.

"78% match consensus on a 1k pilot."
Falsifiable. Could be tested - but no pre-committed kill.

"Pre-committed kill at sub-70% on a held-out set."
Load-bearing under attack. The decision is real.

Same claim. Different evidence. Different verdict.

Bite.

Are you learning anything? Or restating common wisdom?

What it sounds like at each level:

"AI will help with tagging."
Zero update for an informed reader. Common wisdom.

"LLM tags are more consistent than hand tags."
Plausible. Mild update if pushed; not contrarian.

"Our humans disagree more than LLM-vs-human."
Kappa = 0.34 between our 4 ops. Counterintuitive.

Common wisdom is zero information.

Bet.

Who would bet money on it? Who would sign their name?

How sharp is the dollar claim?

"Improve search experience."
No metric, no magnitude, no signer. Bad proxy.

"1-3% lift on search CVR = $84K-$252K/yr."
Range exists but no one's signed for it yet.

"$200K/yr in redeployed ops capacity."
The VP of Merchandising will sign for the headcount.

Unsigned numbers don't ship.

Paste this into Claude.

Abbreviated for the slide. Full prompt linked in comments + at templeton.host/tools/good-dog

```
You are running Andrew Templeton's DoG Test.
Full version: templeton.host/tools/good-dog

Refuse to evaluate my idea until I have answered:

1. BACKING
   (a) the specific claim
   (b) the evidence behind it
   (c) an observation that would change my mind
2. BITE
   (a) the non-obvious belief this updates
3. BET
   (a) what metric this moves, in dollars
   (b) who would bet money on it / sign their name
```

Five tests. Predictions before runs. Three illustrated below.

n = 5

Each test pairs a starting brief with a synthetic pocket the operator can read via tool use. The subject prompt is fixed; only the brief + pocket vary.

Hand-curated cases, not a factorial matrix - five worked examples covering distinct verdict shapes.

Vision model for shipment manifest QA

Rich pocket. Operator walked in sharp. Audit tightened the definition further.

	Before	Trained DoG
Backing	*"Deflect 60-70% of first-pass review."*	*71% recall @ 99% precision; kill criteria named.*
Bite	*"Vision models are mature enough now."*	*70% auto-pass vs peer-published 20-40% ceiling.*
Bet	*"Eight-figure cost surface."*	*$3.5-4M/yr net; CFO + Ops Director signed.*

GOOD DoG

Audit didn't change the verdict - it tightened the definition until it could be signed.

LLM-generated product tags

Same pitch as Case 1 - different pocket data. Audit surfaced a different project entirely.

	Before	Trained DoG
Backing	 *"LLM tags are more comprehensive."*	 *78% consensus on pilot; humans kappa=0.34.*
Bite	 *"LLMs will improve search CVR."*	 *47% search sessions zero-result; tags miss the bottleneck.*
Bet	 *"Improve search experience."*	 *Real $ - but on a different project (GMV-owner assignment).*

GOOD DoG - for a different project

The audit landed at GREEN, but for the project the dialogue surfaced underneath.

AI chatbot for the support page

Sparse pocket. Operator engaged honestly. Audit graduated to "go gather X first."

	Before	Trained DoG
Backing	*"Vendor demo: 42% deflection."*	*Numeric, but reverse-engineered from vendor marketing.*
Bite	*"AI chatbots help support."*	*Real bottleneck is content ops, not vendor choice.*
Bet	*"Improve CX."*	*Dollar ceiling priced; no signer till categories measured.*

NEEDS TRAINING

The project that survives this audit is not the project the operator pitched.

Every case showed the same move.

Before the audit: the pitch could mean anything.
After the audit: the pitch is one specific thing someone can commit to.

Either everyone aligns on the same idea -
or it gets exposed as too vague to evaluate.

Align - or it's just bad.

Run the test.
You get one of three answers.

GOOD DoG
Everyone aligns on why the idea is good. You can ship.

NEEDS TRAINING
You learn what further data the idea needs before it can be evaluated.

BAD DoG
The idea is exposed as bad, before more time gets spent on it.

Every answer beats the meeting that produced the vague pitch in the first place.

Train the team to structure bets on good dogs. Not bad ones.

How many bad dogs have bitten you?

Roadmap items shipped without a verdict on what 'good' meant.
The DoG Test catches them while they're still on the page.

Get better at better.

1. **Run the DoG Test:** `templeton.host/tools/good-dog`
2. **Subscribe for more like this:** `templeton.host`
3. **Follow on LinkedIn:** `/in/andrewtempleton`

More AI techniques like this monthly.