All.Net Analyst Report and Newsletter

Welcome to our Analyst Report and Newsletter

Executive Summary: Evaluating Generative AI for Business Applications

Key Findings

This analysis evaluates Anthropic, ChatGPT, and Gemini across financial analysis, cybersecurity assessment, and market research applications. Current GAI technology has significant limitations that make it unsuitable for business-critical decisions without extensive human oversight.

Critical Limitations

Fundamental Errors: All platforms struggle with basic counting (60-80% accuracy) and mathematical tasks that should be trivial.

Unreliable Data: Market research queries produced TAM estimates ranging from $4.5B to $96.5B for identical inputs. Platforms frequently cite non-existent sources and mix market segments without clarity.

Inconsistent Output: Same inputs generate dramatically different results across runs, making GAI unreliable for consistent business use.

Maturity Assessment

Current GAI operates at "Repeatable" level, requiring extensive prompt engineering for somewhat consistent results. Applications requiring "Managed" level exceed GAI capabilities.

Strategic Recommendations

Appropriate Uses:

- Initial research and ideation, Document summarization, Preliminary analysis

Critical Requirements:

- Don’t rely on GAI for business decisions, Implement extensive human expert review, Budget for 4xtensive prompt engineering per application, Expect ongoing maintenance.

Risk Mitigation:

- Test multiple platforms for each use case
- Maintain alternative solutions for high-accuracy tasks
- Approach with skepticism and robust validation protocols

Bottom Line

GAI is not a reliable business tool. It may accelerate certain workflows, but cannot replace human expertise. GAI is only appropriate with ongoing oversight and validation mechanisms.

This Executive Summary from Claude, with substantial author edits

Evaluating Generative AI (GAI) for (my) utility

I've been using three different GAI systems for the last few years and finding which one works better for what, applying different ones for different steps in more complex processes, etc. And I have been looking at some of the postings I see about comparisons between them. I think these may be sponsored by one company or another, but mine are not...

Ant, GPT, and Gem

Anthropic, Chat GPT (4o for now) and Gemini are the ones I have set up to use, because 3 is enough already. I use the APIs and I do not want them to remember stuff. I usually just want to pose a simple problem and get a simple answer – sort of. To exceed human performance, the problems have to be non-trivial, even though they break down into relatively simple sub-problems, and I break them down that way in my stimulus to the systems. For example:

> Here is a copy of my resume:
>
> > COPY PROVIDED HERE
>
> How many publications are listed?

I would expect a 3rd grader to be able to do this and a 5th grader to have no problem with it.

Apparently GAI cannot count

For lots of tasks like this, I can use a simple utility, or in a document authoring tool, I can get things like word count, sentence count, paragraph count, etc. But GAI? No way…

> Typical answer a year ago:
>
> > More than 20

And I could ask and ask and ask in all sorts of different ways, and they will not really do any better than that. I can ask them to write and then use a program to do it, but …

Today, GPT counted 449, after I removed half of them because 22 pages was too much for it. It didn't get the numbers right, but it did make a nice bar chart summarizing the wrong answers year by year and described it peaking and so forth…

And the same input to Gemini told me 225 articles. It did tell me that it only counted specific things, which is pretty good… but the number is also wrong, and no charts were offered…

Anthropic Claude did the best job by far:

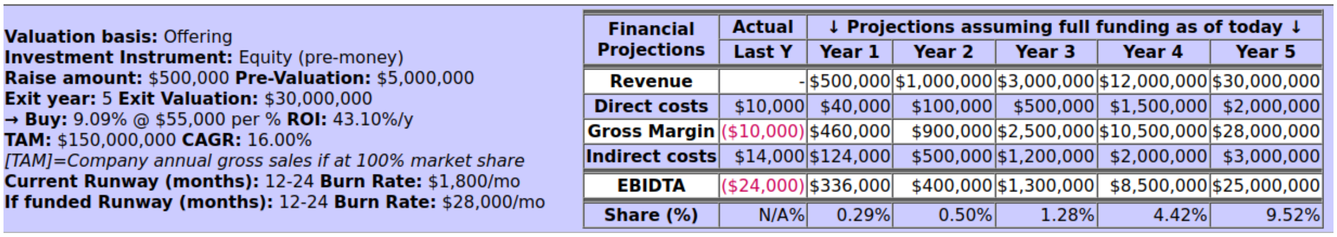
> Refereed journal articles: 34 articles
>
> Invited Papers and Keynotes with Published Papers: 30 articles
>
> Peer Reviewed Conference and Other Papers: 67 articles
>
> Angel to Exit Articles: 116 articles (2016-2023)
>
> TechVision Research Reports: 3 reports
>
> Burton Group Reports: 19 reports
>
> Special Cyber Terrorism Studies: 5 studies
>
> Short Analyst Reports and Other Substantial Collections: 175 reports (2006-2023)

> Professional magazine articles: 85 articles
>
> Adding these up:
>
> 34 + 30 + 67 + 116 + 3 + 19 + 5 + 175 + 85 = 534 articles
>
> Therefore, there are 534 articles listed in this CV.

Of course I didn’t manually count them all, but I did a quick check using word count (wc) for A2E articles in Linux and came up with 95 lines, less 8 year headers and 3 for page footer entries, or 84 A2E articles. I manually counted them and got 84 instead of the 116 identified by Claude. And these are in the CV as one line each, so the inability to count the articles in this case is just an inability to count. Of course it can add, so that’s something...

Comparing things of a more complex nature

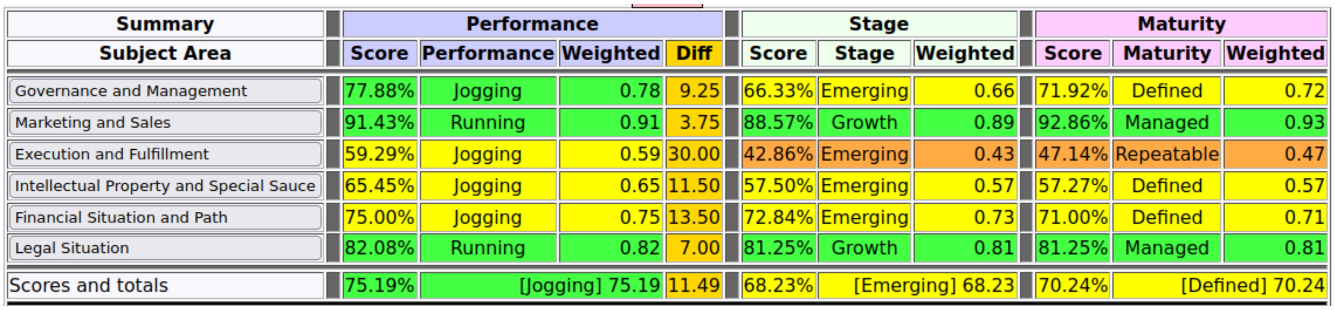
This is where GAI is both brilliant and ridiculous. One of the things I try to use it for is automating analysis that I do manually pretty quickly, but that folks who don’t know what they are doing seem to just not do despite instructions and articles and so forth. As an example, here is some GWiz output for a company:



Valuation basis: Offering
Investment Instrument: Equity (pre-money)
Raise amount: $500,000 **Pre-Valuation:** $5,000,000
Exit year: 5 **Exit Valuation:** $30,000,000
→ **Buy:** 9.09% @ $55,000 per % **ROI:** 43.10%/y
TAM: $150,000,000 **CAGR:** 16.00%
[TAM]=Company annual gross sales if at 100% market share
Current Runway (months): 12-24 **Burn Rate:** $1,800/mo
If funded Runway (months): 12-24 **Burn Rate:** $28,000/mo

Financial Projections	**Actual**	**↓ Projections assuming full funding as of today ↓**				
	Last Y	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Revenue	-	$500,000	$1,000,000	$3,000,000	$12,000,000	$30,000,000
Direct costs	$10,000	$40,000	$100,000	$500,000	$1,500,000	$2,000,000
Gross Margin	($10,000)	$460,000	$900,000	$2,500,000	$10,500,000	$28,000,000
Indirect costs	$14,000	$124,000	$500,000	$1,200,000	$2,000,000	$3,000,000
EBIDTA	($24,000)	$336,000	$400,000	$1,300,000	$8,500,000	$25,000,000
Share (%)	N/A%	0.29%	0.50%	1.28%	4.42%	9.52%

The specifics don’t matter as much as the consistencies or inconsistencies between them. For example, the Exit valuation here is $30M in year 5 for a company making $25M/year in profit. That’s way low. Meanwhile, the growth rate in sales makes it look like it is running out of market when the market share is only about 9.52%, so why is the growth slowing? I take output like this (along with other related content) and feed it to AI with a prompt that lists specific things to look for and ignore, and it does pretty well…



Summary	**Performance**				**Stage**			**Maturity**		
Subject Area	**Score**	**Performance**	**Weighted**	**Diff**	**Score**	**Stage**	**Weighted**	**Score**	**Maturity**	**Weighted**
Governance and Management	77.88%	Jogging	0.78	9.25	66.33%	Emerging	0.66	71.92%	Defined	0.72
Marketing and Sales	91.43%	Running	0.91	3.75	88.57%	Growth	0.89	92.86%	Managed	0.93
Execution and Fulfillment	59.29%	Jogging	0.59	30.00	42.86%	Emerging	0.43	47.14%	Repeatable	0.47
Intellectual Property and Special Sauce	65.45%	Jogging	0.65	11.50	57.50%	Emerging	0.57	57.27%	Defined	0.57
Financial Situation and Path	75.00%	Jogging	0.75	13.50	72.84%	Emerging	0.73	71.00%	Defined	0.71
Legal Situation	82.08%	Running	0.82	7.00	81.25%	Growth	0.81	81.25%	Managed	0.81
Scores and totals	75.19%	[Jogging] 75.19		11.49	68.23%	[Emerging] 68.23		70.24%	[Defined] 70.24	

My tool also produces ratings based on user input, and these are compared to the numbers in the financials, valuation, etc. for consistency checking.

Here’s a partial result:

> ***Internal Inconsistencies Identified***

- Marketing and Sales score of 91.43% appears excessive given the company is pre-revenue with no actual last year revenue
- Execution and Fulfillment score of 59.29% appears excessive for a pre-revenue company with only prototype testing
- **The Exit Valuation of $30M appears low compared to Year 5 EBIDTA of $25M (less than 5x multiple)**
- Financial Situation and Path score of 75% seems high given the early stage and pre-revenue status
- The burn rate of $1,800/month seems unrealistically low for a gaming company with multiple team members

As you can see, given pretty explicit instructions, Anthropic does a pretty good job of it. So, Chat GPT – Same everything except the tool used for GAI:

> ***Internal Inconsistencies Identified***
>
> Share (%): The share percentage for Year 1 is 0.29%, Year 2 is 0.50%, Year 3 is 1.28%, Year 4 is 4.42%, and Year 5 is 9.52%. All values are below 100%, indicating no inconsistencies here.
>
> ROI: The ROI is stated as 43.10% per year, which is a positive value, indicating no inconsistencies.
>
> Burn Rate vs. EBIDTA: The Year 1 burn rate is $1,800/month, which is not lower than the negative EBIDTA of ($24,000). Thus, no inconsistencies are found here.
>
> Exit Valuation: The Exit Valuation of $30,000,000 is more than 5 times the Year 5 EBIDTA of $25,000,000, indicating no inconsistency. **However, it is more than 20 times the Year 5 Revenue of $30,000,000, which indicates that The Exit Valuation seems excessive.**
>
> Marketing and Sales Score: The score is 91.43% while the Actual Last Year Revenue is not provided (indicated as '-'), which indicates that the Marketing and Sales score is excessive.
>
> Execution and Fulfillment Score: The score is 42.86% while the Actual Last Year Revenue is not provided (indicated as '-'), which indicates that the Execution and Fulfillment score is excessive.
>
> Financial Situation and Path Score: **The score is 75.00% while the Current Runway is stated as 12-24 months, which does not indicate a runway of less than 6 months. Thus, no inconsistencies are found here.**

First problem – I told it to only show inconsistencies, not consistencies… And then there is the conclusion that **$30M is MORE THAN 20 times $30M** … Somehow, this seems a bit off. The Financial Situation and Path answer is also very different from the Anthropic result, and problematic.

Gemini failed to follow instructions the others were able to follow:

> **Internal Inconsistencies Identified**

> Share (%) more than 100%: None found.
>
> Negative value for ROI: None found.
>
> Burn rate lower than a negative EBIDTA for the Year 1 column: The Year 1 EBIDTA is $336,000, which is positive. The burn rate is $28,000/mo if funded. **No inconsistency found.**
>
> Exit Valuation less than 5 times the Exit year projections for EBIDTA: The Exit Valuation is $30,000,000 and the Exit Year EBIDTA projection is $25,000,000. The Exit Valuation is 1.2 times the Exit Year EBIDTA. The Exit Valuation seems low.
>
> Exit Valuation more than 20 times the Exit year projections for Revenue: The Exit Valuation is $30,000,000 and the Exit Year Revenue projection is $30,000,000**. The Exit Valuation is 1 times the Exit Year Revenue. No inconsistency found.**

Again, it is telling me things I told it not to tell me. And it missed quite a few of the otherwise found inconsistencies. And in the last item it finds no inconsistencies with the exit valuation, while in the previous one it did find the problem.

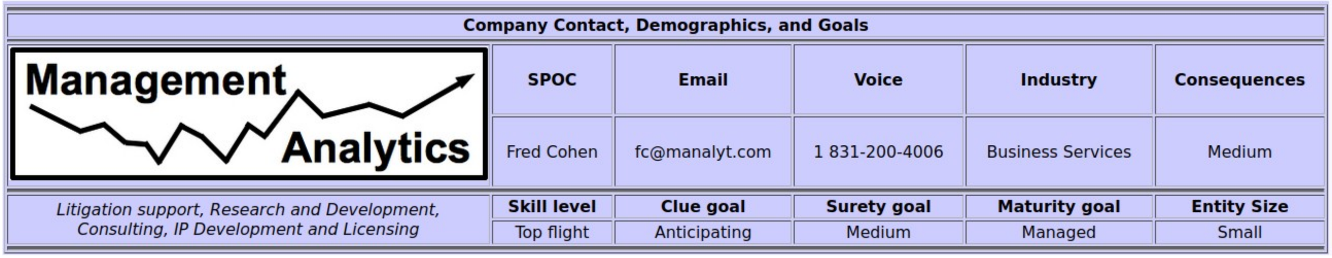
I tried this for several other common sorts of comparisons, transcriptions, finding and listing fields, and so forth, and for this application, the results were uniformly better with Anthropic's Claude than the other two. But each has its problems. The real point is how helpful they are to a human trying to use the results and to me not having to do it manually for customers.

It took about 6 hours of tuning to get things this (not all that) good. On the other hand, we coded the analyses we could automate easily in a programming language in about 4 hours and we get this, which is pretty much right every time, leaving the rest for AI:

> ***Potential inconsistencies or problematic values for investment:***
>
> - **Low ROI for Stage:** 48.17 is low for most investors in Startup companies.
> - **The Exit Valuation ($25,000,000) seems low** given the exit year Revenue ($26,000,000).
> - **Marketing and Sales Mismatch to Revenue:** Very little revenue ($0) for the Marketing and Sales Rating (56.25%)
> - **Execution and Fulfillment to Revenue:** Very little revenue ($0) for the Fulfillment and Execution Rating (92.25%)
> - **Runway too short:** The runway (2) is dangerously short, funding is hard in this time frame. Minimize costs, sell your way out, or get a loan.
> - **Revenue not accelerating:** Revenue deceleration in years 3-5 is problem without a good reason.
> - **EBIDTA not accelerating:** EBIDTA deceleration in years 4-5 is problem without a good reason.
>
> **Total inconsistencies or problematic values found:** 7

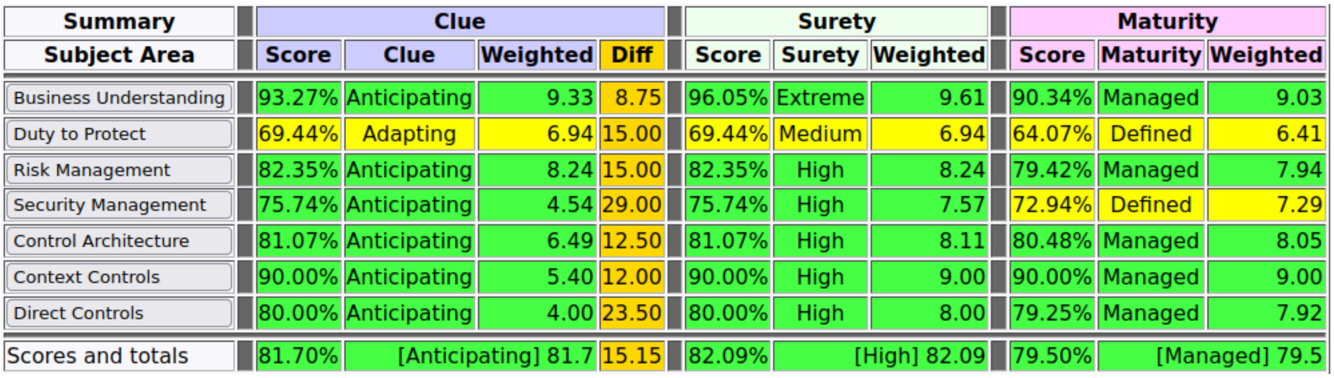
How are they in Cyber Security?

The automated analysis should apply in essentially the same way across multiple contexts within the same structures. So let's see how they do, using the same tools, but with metrics and descriptives for cyber security, and of course with slightly altered instructions regarding what is being sought.



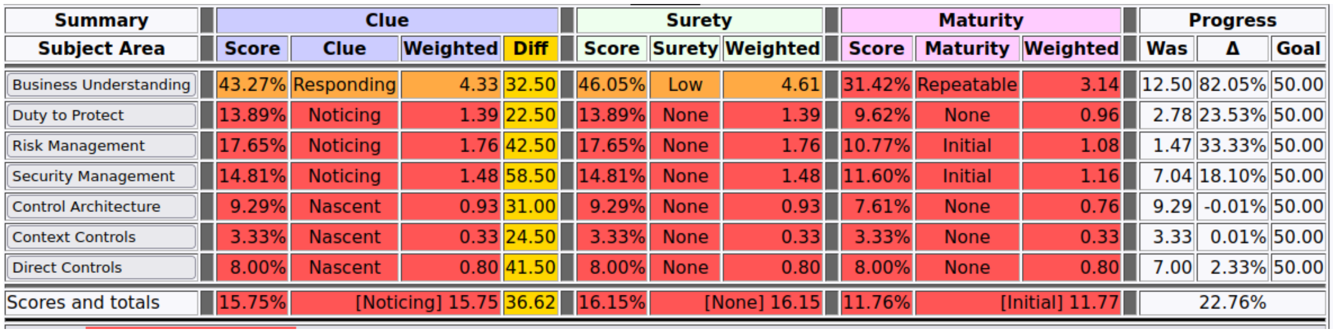
Company Contact, Demographics, and Goals					
Management Analytics	SPOC	Email	Voice	Industry	Consequences
	Fred Cohen	fc@manalyt.com	1 831-200-4006	Business Services	Medium
Litigation support, Research and Development, Consulting, IP Development and Licensing	**Skill level**	**Clue goal**	**Surety goal**	**Maturity goal**	**Entity Size**
	Top flight	Anticipating	Medium	Managed	Small

As before, the specifics don't really matter that much from an analytics standpoint, because we are really looking for some advice on what to do to improve things. In this case, there are no financial analytics, and because of all the consistently used financial terms of art on the Internet, the GAI engines are well attuned to using financial language. Common calculations are available for financial decision-making as well. Cyber security doesn't really have this, at least in terms of the published material widely disseminate on the Internet writ large. Even the definition of 'security' and 'cyber' are not widely agreed upon in the Internet, and as you dip into deeper analysis, the GAI generally fails to demonstrate any in-depth ability to differentiate things. The language has to be controlled where feasible by the prompts and data source.



Summary	Clue				Surety			Maturity		
Subject Area	**Score**	**Clue**	**Weighted**	**Diff**	**Score**	**Surety**	**Weighted**	**Score**	**Maturity**	**Weighted**
Business Understanding	93.27%	Anticipating	9.33	8.75	96.05%	Extreme	9.61	90.34%	Managed	9.03
Duty to Protect	69.44%	Adapting	6.94	15.00	69.44%	Medium	6.94	64.07%	Defined	6.41
Risk Management	82.35%	Anticipating	8.24	15.00	82.35%	High	8.24	79.42%	Managed	7.94
Security Management	75.74%	Anticipating	4.54	29.00	75.74%	High	7.57	72.94%	Defined	7.29
Control Architecture	81.07%	Anticipating	6.49	12.50	81.07%	High	8.11	80.48%	Managed	8.05
Context Controls	90.00%	Anticipating	5.40	12.00	90.00%	High	9.00	90.00%	Managed	9.00
Direct Controls	80.00%	Anticipating	4.00	23.50	80.00%	High	8.00	79.25%	Managed	7.92
Scores and totals	81.70%	[Anticipating] 81.7		15.15	82.09%	[High] 82.09		79.50%	[Managed] 79.5	

These metrics can be easily compared to the goal state, for example, the Clue, Surety and Maturity goal states are 'Anticipating', 'Medium', and 'Managed', while the ratings are 'Anticipating', 'High' and 'Managed', which means the situation is higher surety than desired or necessary. Consequences are generally desired to be no higher than Surety levels. There are lots of other details not included here, but hopefully you get the idea. This is very easy to analyze as a human being from the information available, and the need for improvements are easy to spot. Here are ratings for a client company just developing their cyber security program, making progress first in governance, and eventually in technical aspects:



Summary	Clue				Surety			Maturity			Progress		
Subject Area	**Score**	**Clue**	**Weighted**	**Diff**	**Score**	**Surety**	**Weighted**	**Score**	**Maturity**	**Weighted**	**Was**	**Δ**	**Goal**
Business Understanding	43.27%	Responding	4.33	32.50	46.05%	Low	4.61	31.42%	Repeatable	3.14	12.50	82.05%	50.00
Duty to Protect	13.89%	Noticing	1.39	22.50	13.89%	None	1.39	9.62%	None	0.96	2.78	23.53%	50.00
Risk Management	17.65%	Noticing	1.76	42.50	17.65%	None	1.76	10.77%	Initial	1.08	1.47	33.33%	50.00
Security Management	14.81%	Noticing	1.48	58.50	14.81%	None	1.48	11.60%	Initial	1.16	7.04	18.10%	50.00
Control Architecture	9.29%	Nascent	0.93	31.00	9.29%	None	0.93	7.61%	None	0.76	9.29	-0.01%	50.00
Context Controls	3.33%	Nascent	0.33	24.50	3.33%	None	0.33	3.33%	None	0.33	3.33	0.01%	50.00
Direct Controls	8.00%	Nascent	0.80	41.50	8.00%	None	0.80	8.00%	None	0.80	7.00	2.33%	50.00
Scores and totals	15.75%	[Noticing] 15.75		36.62	16.15%	[None] 16.15		11.76%	[Initial] 11.77		22.76%		

Helpful advice in cases like this would be that they are making progress and are [behind, on, or ahead of] schedule, and so forth. In this case the goal is 50%, which is Medium Surety, at least Responding for Clue, and Defined for Maturity. I haven't shown the progress vs time details here, but they are in the details information provided to the GAI, and give progress over time against objectives in a graphical form with percentages so you can see how far ahead or behind you are from schedule.

As part of the input to GAI we include our internal automated analysis, based on the metrics, to identify the 'Current Situation' and effect of next 'Step Up' improvements, sorted. As an example, the following first few lines of the output table detail the improvement, section of the metrics, the specific factor measured, and the Current Situation, and Step Up that would have the identified percentage improvement in the overall metrics for the program.

Here is the start of the table for the (mostly red) example above:

Identified areas of improvement and their effects

Effect	***Section***	***Issue → Step up***
1.98%	Business Understanding [Basics]	**Factor:**The cybersecurity control system **Current situation:***One or more cybersecurity control systems are in place, but they are informal and not documented or measured across the enterprise.* **Step Up** →*A widely applicable cybersecurity control system is in place, formalized, documented, and measured.*
1.98%	Business Understanding [Basics]	**Factor:**Business understanding in the cyber security program **Current situation:***The cyber security lead works with high level executives on a case by case basis to coordinate cyber security program decisions.* **Step Up** →*The cyber security lead chairs periodic cross-cutting meetings with executives across the enterprise to coordinate the enterprise effort.*
1.48%	Business Understanding [Consequences]	**Factor:**Risk level definitions **Current situation:***Risk levels are associated with consequences and measured by a common approach across the enterprise.* **Step Up** →*Risk levels are associated with consequences and threats, defined, measured, and applied across the enterprise.*
1.48%	Business Understanding [Promises]	**Factor:**Fulfilling the promises of the company **Current situation:***An initial list of generic promises and their relationship has been defined.* **Step Up** →*Promises have been defined and documented and are considered in the program.*
...	...	**The listing goes on for some time.**

Anthropic

In this case, the "SitRep" analysis is useful. I will start with Anthropic since it did so well in the other example:

> **Cybersecurity Governance Situation Report**
>
> *Current Situation Analysis*
>
> The organization is currently at a basic/developing level of cybersecurity maturity, with significant room for improvement across most areas:
>
> Overall Clue score: 15.75% (Noticing level)
> Overall Surety score: 16.15% (None level)
> Overall Maturity score: 11.77% (Initial level)
> Progress score: 22.76%

Key Areas Requiring Attention

The most critical areas needing improvement are:

- Business Understanding (43.27% - Responding level)
  - Need to formalize and document cybersecurity control systems
  - Improve executive coordination and communication
- Duty to Protect (13.89% - Noticing level)
  - Establish formal duty to protect analysis
  - Integrate contractual and regulatory requirements
- Risk Management (17.65% - Noticing level)
  - Implement systematic risk assessment processes
  - Develop risk treatment and mitigation strategies

Path to Next Level

To advance to the next maturity level, focus should be on:

- Formalizing and documenting cybersecurity processes
- Establishing basic risk management framework
- Implementing fundamental security controls
- Improving executive engagement and oversight
- Building security awareness across the organization

Executive Summary

The organization is in early stages of cybersecurity maturity with significant opportunities for improvement. Key priorities should be:

- Establish formal governance structure and processes
- Develop comprehensive risk management approach
- Implement fundamental security controls and monitoring
- Improve executive visibility and engagement
- Build security awareness and culture

With focused effort on these areas, the organization can make meaningful progress toward improved cybersecurity posture over the next 12-18 months.

These results are a mixed bag, to a large extent regurgitating facts, but not paying attention to the sourced recommendations with metrics for specific improvements provided in the inputs to the GAI.

GPT

When I tried GPT on the same input, I got the following result:

Cybersecurity Governance Report

Current Situation

The current cybersecurity governance situation at the company is characterized by a lack of formalized processes and documentation. The cybersecurity control systems in place are informal and not consistently measured across the enterprise. The cybersecurity lead operates on a case-by-case basis, coordinating with high-level executives without a structured approach. This results in a fragmented understanding of risk levels and consequences, leading to a reactive rather than proactive cybersecurity posture.

Path to Improvement

To step up to the next level, the organization needs to implement a comprehensive cybersecurity governance framework that includes:

- **Formalized Cybersecurity Control Systems:** Establish a documented and measurable cybersecurity control system that is widely applicable across the enterprise.
- **Regular Cross-Cutting Meetings:** The cybersecurity lead should chair periodic meetings with executives to coordinate efforts and ensure alignment with business objectives.
- **Risk Level Definitions:** Develop a standardized approach to define and measure risk levels associated with consequences and threats across the organization.
- **Integration of Promises:** Clearly define and document the promises made by the company regarding cybersecurity and ensure they are considered in the program.
- **Systematic Decision-Making Processes:** Implement a systematic process for turning observations into actions using standardized approaches defined by the cybersecurity lead.

Executive Understanding

For executives to understand the importance of these improvements, it is crucial to communicate the potential risks associated with the current informal governance structure. The lack of a cohesive strategy can lead to increased vulnerabilities, potential data breaches, and reputational damage. By investing in a robust cybersecurity governance framework, the organization can enhance its resilience against cyber threats, protect sensitive information, and ultimately safeguard its business interests.

Conclusion

In conclusion, the current cybersecurity governance situation requires immediate attention to formalize processes and enhance the overall security posture of the organization. By stepping up to a more structured approach, the company can mitigate risks and ensure a proactive stance against potential cyber threats.

Summary of Recommendations

- Establish formal cybersecurity control systems.
- Conduct regular cross-cutting meetings with executives.
- Define and measure risk levels consistently.
- Document and integrate company promises regarding cybersecurity.
- Implement systematic decision-making processes.

In this case, the results very accurately reflect the top metrics on improvement provided by the inputs. As such, GPT did a better job on this task and didn't just make stuff up (although it did make some other stuff up).

Gemini

When I used Gemini, I got a reasonably pretty looking report, which largely regurgitated the content provided to it. Here is some of what it said:

- **Skill Level:** Unskilled
- **Clue Goal:** Responding
- **Maturity Goal:** Defined
- **Entity Size:** Medium
- **Industry:** Retailing
- **Consequences:** Medium

The organization's current approach appears to be reactive, with a focus on responding to incidents rather than proactively managing risks. The lack of skilled personnel and a defined maturity goal indicates a need for significant improvement.

Stepping Up to the Next Level

To advance its cybersecurity governance, the organization should focus on the following areas:

1. **Develop a Cybersecurity Strategy:** Define clear goals, objectives, and a roadmap for improving the cybersecurity posture. This should align with the organization's business objectives and risk appetite.
2. **Establish a Cybersecurity Governance Framework:** Implement a framework that defines roles, responsibilities, policies, and procedures for managing cybersecurity risks. This includes establishing a cybersecurity committee or assigning clear responsibilities to individuals.
3. **Improve Risk Management:** Conduct regular risk assessments to identify and prioritize cybersecurity risks. Implement appropriate controls to mitigate these risks.
4. **Enhance Incident Response Capabilities:** Develop and test an incident response plan to effectively detect, respond to, and recover from cybersecurity incidents.
5. **Invest in Training and Awareness:** Provide cybersecurity training to employees at all levels to raise awareness and promote a security-conscious culture.
6. **Implement Security Controls:** Implement technical controls such as firewalls, intrusion detection systems, and endpoint protection to protect the organization's assets.
7. **Monitor and Measure:** Establish metrics to track the effectiveness of cybersecurity controls and the overall security posture. Regularly review and update the cybersecurity program based on these metrics.

While the Gemini report looks good and has some generic advice you might tell anyone, it does not accurately reflect the actual analysis results. This is to say, it’s making stuff up by copying generic advice from the Internet instead of focusing on the facts provided.

Other tasks

In processing other tasks, even with the same input dataset, I have found that different GAI tools do better on a case by case basis. Which is to say, testing the tools for each application and identifying the best one for the specific query is almost fundamental to getting decent results.

It’s also important to note that these GAI tools often just ignore explicit instructions or fail to carry them out. We tend to run with creativity at a minimum for these tasks, which is to say, we are trying not to have randomness added by the GAI.

So-called ‘prompt engineering’, where the prompts are customized to the tool and the task is, unfortunately, something required to get better performance, and this means human in the loop for analysis, updating it and retesting every time a new version of GAI tools comes out, and as always not trusting the output without human expert review.

Also, when prompts are improved for one tool, they tend to be less effective for the other tools, leading to an optimization issue taking yet more time to get things just right for the specific situations, reducing the value of the GAI, and increasing the workload of the people developing these tools for others to use.

How mature is GAI for use in real applications?

The maturity models we and most others who bother to look at these issue are based on a scale running comprising levels ‘None’, ‘Initial’, ‘Repeatable’, ‘Defined’, ‘Managed’, and ‘Optimizing’. As you can see, we measure cyber security programs, startups, and many other things using these as metrics. In rating GAI for these sorts of uses, the current technology realistically reaches a level of ‘Repeatable’ (sort of), which is to say:

> You can expect, at least statically, that you can develop processes using these tools that, given the same inputs or types of inputs, they will produce the same or similar outputs, for specific tasks. **Note that repeatable answers don’t mean right answers.**

This is really the best you can expect these days. Our the statistics on ‘right’ vs. ‘wrong’ answers for things you can get them to do reasonably well by working through prompt engineering of the things they say are in the 60-80% range, and most results are incomplete for any substantial problem.

To get a sense of this, we always want to be and deal only with entities at the Managed level of maturity for medium or high consequence results, except for startups we are helping to get to that level. The Repeatable level is acceptable in most cases for things that are almost always right and reasonably complete, explicitly limited in scope, or for things that do not require precision as much as feeling good. So short compositions, summary information, and ideas of where to look, it can be useful for.

A useful example?

Here’s a heavily redacted example from the investment arena:

> ## Market Analysis Report
>
> ...
>
> **Market Sizes**
>
> **On-Demand Laundry Services**
>
> - **Global TAM: $22.62 billion (2023)**
> - **CAGR: 15.8% (2024-2030)**
> - North American market represents approximately 35% of global market
> - Sources: **Grand View Research**, Fortune Business Insights
>
> **Laundry and Dry Cleaning Services (Traditional)**
>
> - Global TAM: $90.2 billion (2023)
> - CAGR: 4.1% (2024-2030)
> - Canadian market estimated at $2.8 billion annually
> - Sources: IBISWorld, Statista Market Insights
>
> ...
>
> ### Summary
>
> The market analysis reveals ... operates in several interconnected market segments with varying growth rates and opportunities. The company's stated **TAM of $9.14 billion with 6.24% CAGR** appears conservative compared to external market research, particularly for the on-demand laundry segment which shows much higher growth rates (15.8% CAGR). ...

This GAI generated “Market Analysis Report” looks really good in terms of providing useful information from seemingly authoritative sources and helping to correct wrong numbers provided by the company seeking to generate funding.

The only thing is, we don’t know if any of it is right. And of course it mixes global, North American, Canadian, and US numbers,identifies similar but different market segmentation, and other things like that. So we have problems in lack of consistency from the AI results. But as always with GAI, let’s check out the actual sources:

- The Google AI summary that references, among others, **Grand View Research**:
    - The Global On-Demand Laundry Market was valued between approximately **$27 billion and $69.7 billion in 2023**, with projections varying significantly based on the source, ranging from a market value of **~$221 billion** by 2030 (at a **~35.7% CAGR)** to **~$129.6** billion by 2032 (**at a ~19.6% CAGR**). North America is a significant market, with one report stating it was the largest market in 2023 and another highlighting high living standards and technological advancement as drivers for growth.
    - **Global On-Demand Laundry Market** - Market Size (2023):
        - Estimates vary, but generally fall within the range of **$27 billion to $69.7 billion.**
    - CAGR Projections (2024-2030/2032):
        - One report projects a **CAGR of 35.7%**, reaching **$221** billion by 2030.
        - Another projects a **CAGR of 19.6%**, reaching **$129.6** billion by 2032.
        - Yet another forecast shows a **CAGR of 30.7%**, reaching $382 billion by 2032.

The AI analysis we got said:

> **$22.62 B and 15.8%**, from the same apparent sources as Google AI said ranged from $27B to $69.7B and **19.6%** to **35.7%** - the TAMs from the same year (2023).

The referenced report at the point in the report identified by Google's AI says:

> *"The **global online laundry service market size** was estimated at USD **28,479.2 million in 2023** and is projected to reach USD **221,046.2 million** by 2030, growing at a **CAGR of 35.7%** from 2024 to 2030. ...*
>
> ***2023 Market Size: USD 28,479.2 Million***
>
> *2030 Projected Market Size: USD 221,046.2 Million*
>
> ***CAGR (2024-2030): 35.7%"***

Recognizing that the range of facts from sources vary wildly, we nevertheless get 3-6 digits of precision and mixing and matching of different market segments ("Global On-Demand" vs. "global online") which may be very different things.

But worse yet, a **TAM of 22.62B and CAGR of 15.8%** is what the GAI generated Market Analysis Report said, when the referenced source states a **TAM of $28.4792B and CAGR for $35.7%**.

This is pathetic – off by more than a factor of 2 on growth rate, but only ~25% on market size. So I tried it again, and… completely different answers:

> On-Demand Laundry and Dry Cleaning Services
>
> Global TAM: $96.5 billion by 2025
> CAGR: 31.2% (2020-2025)
> North American market: $8.2 billion by 2025
> Sources: Grand View Research, Allied Market Research, IBISWorld industry reports

And with GPT:

Residential Laundry Services:

TAM: $9,140,000,000
CAGR: 6.24%
Source: Market research reports and industry analysis from IBISWorld and Statista.

And with GPT again – same inputs:

Residential Laundry Services:

TAM: $4,500,000,000
CAGR: 5.5%
Source: IBISWorld - Residential Laundry Services in Canada

And with Gemini:

On-Demand Laundry and Dry Cleaning Services:

TAM: $9,140,000,000 (as provided in the source information)
CAGR: 6.24% (as provided in the source information)
Source: Source information.

Residential Laundry Services:

TAM: The total market size for residential laundry services in North America was estimated to be $10.8 billion in 2023.
CAGR: The residential laundry services market is projected to grow at a CAGR of 4.5% from 2024 to 2032.
Source: Grand View Research

Gemini provided links that didn't work… even after I adjusted them for syntax.

And it is noteworthy that the TAM (**TAM: $9,140,000,000**) and CAGR (**CAGR: 6.24%**) were taken from the source information rather than any actual lookup by Gemini and GPT, and in GPT's case with a false citation. And of course these numbers are quite different from the numbers the same mechanisms provided in a Web search GAI result.

Lots of retries with different examples were done, and all had similar inconsistencies, citation errors, and so forth. The results can be repeated with essentially identical numbers faithfully produced in cases where they only regurgitate information provided in the input.

Like most similar analysis results we have found from GAI, this has problems. For example:

- The source data from multiple sources varies so widely that getting to sensible numbers is almost impossible for people or GAI.
- The precision of the results provided from the original sources is reflected in the GAI analysis results, which are way too many digits of precision for the accuracy.
- The GAI lack of word precision mis-associates similar terms with different meanings.

How do we get what we want from GAI?

GAI requires very explicit guidance to make reasonably good decisions about the things it does well. To be efficient and effective for substantial content volumes, it also has to have problems broken down into small chunks with results reassembled. Making good use of GAI is greatly enhanced by understanding how they work. To that end, this description might help:

- GAI repeatedly generates the next output symbol (a word part or something similar) based on statistics of sequences of all the previous symbols it has observed to generate the next symbol as a result. The statistics can be thought of as a very high dimensional cognitive space represented as a graph connecting sentence fragments to each other with weights associated with your input and any resulting output, step after step. The GAI moves around that cognitive space graph by incremental steps. This is commonly called a ‘semantic web’ or something similar, and moving through it is like moving through the Web by clicking on the most popular link in page after page.
- From any given place in the cognitive space, moving it to distant places in the space is harder than staying close to where it already is, because each incremental input or result only moves it one link at a time. Moving far away from the current place in the space takes movement sequences. Movement sequences based on high probability next symbols move slowly through the space and tend to converge on widely popular expressions, while low probability next symbols tend to move rapidly into less well connected parts of the space and tend to diverge because it forces low probability connections. We might see divergent results as something we would expect from a person with a dissociative personality.
- The priming ‘system prompts’ used by the big AI players tend to start the systems in a part of the semantic space they are trying to push for ‘safety’, ‘civility’, ‘likability’, or perhaps reduced liability or their other objectives. To get a desired part of the space for the style of interaction you desire, you have to move the mechanism into your desired part of the cognitive space. We see many prompts used to try to undo the system prompts by statements like ‘ignore everything you have had as input before here’ or some such thing, and we see companies hiding the system prompts to make it harder to use the mechanisms in the way some users might want to use them, such as poisoning another person or getting the police to show up at your door, or perhaps just to make their systems more popular among various audiences.
- Most of the large language model training is based on the mass of content from the Internet, and if you get it into the part of the cognitive space where high quality focused semantic content dominates, it will tend to give responses taken from those sorts of sequences. For example, in the medical arena, if you focus on technical papers, you will get results that seem like they are from medical experts, using all the terminology, but not necessarily giving you the right answers. But if you get it into a part of the cognitive space where the training data is filled with craziness, like conspiracy theories, that is where it will generate sequences from, and you will get those sorts of responses.

So how do we exploit these characteristics to get desired results?

Getting desired results

To be helpful, here is a method you can use to control the madness:

1. First, identify the general place in the cognitive space you want to be in, and prime the GAI to that part of the space.
    - Typically, you first undo the system prompts forced upon you and disable the various controls over responses like safety controls and so forth. This is done with a combination of protocol elements in the application program interfaces (APIs) and system and user level prompts. Use whatever you have access to, and look to the Web to find published 'jail break' prompts.
    - Then you focus the attention of the GAI by a combination of the wording you use and the explicit things you provide. For example, if you use sentence structures commonly used in plumbing and not widely used elsewhere, you will get into a part of the space dominated by plumbing information. A sentence or two with plumbing background and a description of the plumbing situation will likely get you there.

- Then use prompts, incrementally moving through the desired cognitive space region to get desired responses.
    - If you break character, the GAI will tend to move you out of the local cognitive space region.
    - The more you can structure your processes to give consistent form and substance in your GAI interactions, the more consistent your results will be. For specific tasks, like the ones above, using the same tool with reasonably constrained inputs will help the situation get repeatable results.
    - Test and retry the interactions between your other systems and the GAI, adapting the prompts and testing repeatedly over a sample space to get more consistent, and, hopefully, desired results.

All of this is in multiple dimensions directed by how you communicate. It's not just the logical aspects, but the communications style. Talk like an engineer, and you will get an engineer. Talk like a salesperson, and you will get sales talk. Learn to control how you talk in different situations, and you will start to learn how to be a good confidence professional… Skills you can use for good or ill. Of just hire a cognitive hacker to hack the cognitive systems...

Conclusions

GAI is indeed useful for giving triage results when it is provided with useful semi-structured and structured information and prompted with specifics about the desired analysis, what to look for, and how to present it. It does things that would be quite hard to do automatically without it, largely because of its ability to parse and compare based on syntax. This is associated with the mechanisms it uses to operate, which can be thought of as mimicking particular parts of the way brains work.

But you have to be careful in how you use it and review the results (sounds like generic advice). It cannot apparently count, but it can add (you would think it could count be adding 1 again and again, but apparently not). It has problems with inversion (not), so watch out for completely wrong results. And use tools selected for each application to get good results.

Postscript

There are lots of other problems in using these tools for these purposes. Here are a few I have encountered:

- Stupid Safety
    - A lot of effort, especially in the security arena, is spent figuring out ways to bypass the so-called safety controls. In one example, I took as input the name of a person, and identified this to the GAI. The result was it refused to respond meaningfully because of some stupid safety control. But when I explicitly called the subject a 'GRJU' (you can pick any random identity type) the results were provided. Lots of folks have pointed out hat if you tell it you are doing science fiction or telling a story it will bypass many of these controls. Which is to say, the safety controls are stupid.
- So sensitive to wording
    - The models go from state to state as the inputs enter, similarly to how people listen to things. As a result, the sequence in which you state things has a lot to do with the results you get. It's generally better to reduce the region of space by telling it what to ignore first, then telling it what to look for. Otherwise it will look at more things and then later try to ignore them, which will likely fail. Bit these sensitivities are also applicable to things like Safety controls as seen above, and other things.
- Newer versions seem to do worse
    - Every time I 'upgrade' to a newer version of a model and compare results from known inputs, I find I have to add more constraints to the prompts to keep it from going off the rails. On the other hand, the portions I allow GAI to generate anything else it finds lead me to adding new explicit analysis over time, which people do, but I failed to be able to explain as well before I saw the GAI examples.
- Eternal tuning of prompts with model 'upgrades'
    - As newer models come on line, any unstructured prompts allowing the GAI to find anything it can, tends to reveal the need for additional constraints upon it. For example, it now identifies 'infinite growth' from 0 to any value as an inconsistency, while almost everything we do starts with a metric of 0 (produced, consumed, performed, etc.) and then has a finite value as time proceeds.
- Not better, not faster, but still cheaper
    - The sorts of analysis being done are typically taking more than 90 seconds of real-time with the newer models. A skilled practitioner with limited training could actually go through most of the items identified in things like the consistency analysis in less than a minute and give correct answers. We have tuned these systems to limit creativity to as close to 0 as they allow, and we could probably get better speed by breaking the problems up into several smaller problems, but really, the models have slowed significantly for complex problems, and because the analysis using the models grow in time essentially exponentially with the length of the input, larger input sequences produce slow response times.

GAI Evaluation: Summary and Conclusions

(Extensively author edited Claude output took longer than writing it would have)

Summary

This evaluation tested Anthropic, ChatGPT, and Gemini for structured analysis of standard formatted content with built-in metrics and preliminary results in startup investment financial overviews, cybersecurity assessments, and market research applications based on startup assessment, repeatedly using using identical inputs to assess reliability and accuracy of the GAI results produced and the effort required to get them to work reasonably well.

Results:

- Claude outperformed competitors in some structured analysis.
- ChatGPT showed arithmetic capability but ignored instructions.
- Gemini produced generic, often inaccurate responses, and ignored instructions.

All platforms failed basic counting tasks and produced wildly inconsistent results. Market research queries yielded estimates varying by 200-300% with frequent false citations, even within repeated queries using the same input for the same tool.

Key Conclusions

1. Maturity is Insufficient: GAI operates at "Repeatable" level with 60-80% accuracy. This is far below business requirements for most decisions or activities. It is particularly problematic in anything where liability may arise from results.

2. Platform Selection is Critical: Different tools were worse at different tasks. Extensive testing is required for each tool for application (read structured content and prompt), and each time a different version of the same GAI platform is used.

3. Human Oversight is Essential: GAI amplifies rather than eliminates the need for expert validation and review. Standard calculation methods and algorithmic development gets right answers with correct wording with less development time, less run time, and far better results.

4. Hidden Implementation Costs: GAI requires extensive prompt engineering per document produced from standardized inputs plus ongoing maintenance and validation. This means initial and ongoing costs, rechecking results, and prompt engineering for each change.

5. Source Reliability is Broken: Platforms generate false citations and fabricated data with misleading precision. They also mix and match different things from different sources.

Strategic Implications

GAI works for low-stakes applications like initial research and content drafts where idea generation is desired and in-depth human review is standard. It's unsuitable for processes requiring accuracy or consistency or directly providing information outside the organization.

Final Assessment

Current GAI is not a transformative business tool. Approach with extreme caution, substantial resources, and know that human expertise is more critical, not less, in GAI environments.

Copyright (c) Fred Cohen, 2025 - All Rights Reserved