NEGOTIATING THE GOOGLE AI DEAL

Gemini, agents, Google Cloud and the security stack: how Google packages them, where the money sits, and how buyers should respond

A research note for CIOs, CISOs, IT asset managers and procurement leaders who have received a Google proposal that combines Gemini seats, agent platforms, Vertex AI consumption, a Google Cloud commitment and one or more security products in a single commercial package. It is not tied to any single client situation. Written from the buyer's side of the table. Redress Compliance, 2026.

Key takeaways

- Google now sells AI as one committed spend number, not as products. Gemini seats, agent platforms, tokens, media generation, security and infrastructure are all pulled into the same commit so that every line justifies a bigger floor. The buyer's job is to pull them back apart.
- The list price is not the price. Workspace already includes Gemini in the base plan, agent platform seats carry consumption overage on top, and the token rates on the newest models are introductory and reset upward at the end of 2026. Negotiate the unit that actually bills, not the headline.
- The commit is the deal. Committed spend, committed use discounts, Marketplace drawdown and the ramp schedule decide what the AI estate costs over three years. Every product discount Google offers is recovered through the size and rigidity of the commitment it is attached to.
- Security is the cross sell. Mandiant, Google SecOps, Threat Intelligence and Security Command Center are sold into the same agreement, often after an incident. They should be priced on their own evidence, not as a remedy or a commit filler.
- Buy what you can measure. Stage seats to users with proven uplift, forecast tokens from your own telemetry, and keep everything you cannot yet measure on flexible terms. A credible dual vendor position is the single largest source of leverage.

1. WHAT GOOGLE IS SELLING

The Google AI proposal is not one product. It is at least seven billing surfaces, each with its own metric, its own discount logic and its own trap, presented as a single number. The anatomy is consistent across the estates we have seen:

- **Workspace with Gemini included.** The standalone Gemini add ons were retired in 2025 and Gemini was folded into the Business and Enterprise editions, with base prices raised to absorb it. Every user on the plan now pays for AI whether they use it or not. A separate AI Expanded Access add on sells higher usage limits, video and image generation and deep research to the users who exhaust the included quota.
- **Gemini Enterprise, the agent platform.** Sold per seat in Business, Standard, Plus and Frontline editions, with annual commitment pricing below month to month, plus consumption billing beyond the included quota. This is the surface Google positions as the enterprise AI front door: search across your data, prebuilt agents, and the toolkit to build your own. The seat is the visible price; the overage is the real one.
- **Vertex AI and the Gemini API.** Per token, output priced well above input, with reasoning tokens billed as output. The newest Flash generation carries introductory rates through 31 December 2026 and roughly doubles from 1 January 2027. Provisioned throughput, context caching and regional endpoints each carry their own premium or discount.
- **Agents, avatars and media generation.** Agent Builder and the agent development kit bill through Vertex tokens and tool calls. Veo, Imagen, Lyria and the avatar and video presenter features in Vids and the Gemini app are metered per second of video, per image or per generation, at rates that dwarf text tokens. This is the least forecastable line in the deal and the one Google is happiest to leave open.

- **Gemini Code Assist and developer seats.** Per user per month, bundled with a Google Cloud subscription, with API calls billed on top.
- **The Google Cloud commitment.** A multi year committed spend agreement with tiered discounting, plus resource based and flexible committed use discounts on compute, GPUs and TPUs. Marketplace private offers draw down against the commit, including partner led offers, subject to caps. Everything above sits inside this envelope.
- **The security stack.** Google SecOps priced on a committed daily ingest cap with overage in arrears; Mandiant consulting, incident response retainers and Threat Intelligence on custom quotes with no public list; Security Command Center Enterprise on asset or spend based tiers; Chrome Enterprise Premium per user. All of it can be routed through the same commit.

2. WHAT GOOGLE IS ACTUALLY DOING

- **Converting AI enthusiasm into commit size.** The account team is measured on committed growth. AI is the argument for a larger floor, a longer term and a steeper ramp. The product discounts are real; the commitment that funds them is where the margin is recovered.
- **Selling seats to everyone.** Gemini in Workspace is priced across the whole population. The agent platform pitch is the same: all knowledge workers, day one. Google knows adoption will be uneven. The buyer that signs for everyone pays for the sixty percent who never open it.
- **Pricing the trial, not the term.** Introductory token rates, first year seat discounts and AI credits make year one look cheap. Year two arrives with standard rates, expired credits and a commit floor sized to the year one forecast.
- **Leaving consumption open.** Seat overage, media generation and agent tool calls are the lines Google will not cap unless asked. They are also the lines finance cannot forecast, which is why they belong in the contract as ceilings, not as estimates.
- **Using security as the door.** A Mandiant engagement after an incident becomes a SecOps proposal, which becomes a reason to move logs and identity into Google Cloud, which becomes commit. Each step is reasonable on its own; the sequence is the sales motion.
- **Running on the calendar.** Google's fiscal year ends in December and quarters close in March, June, September and December. Proposals, credits and executive escalations cluster at those points. The buyer's calendar should not.

3. THE SURFACES: METRIC, LEVERAGE AND TRAP

Surface	What bills	Where the leverage is	The trap
Workspace with Gemini	Per user per month, whole population, edition based	Edition mix, staging of Plus and Enterprise, price protection on the base rate, right to true down at anniversary	Paying the AI uplift for every user and accepting the base price rise as fixed
Gemini Enterprise agent platform	Per seat by edition plus consumption above quota	Seat count staged to measured adoption, Frontline and Business tiers for light users, overage rate capped or pooled	All seats at Standard or Plus on day one with uncapped consumption
Vertex AI and Gemini API	Per million tokens, input and output, reasoning as output	Token discount tiers, price hold past the 2026 introductory expiry, successor model pricing, caching and batch rates	Forecasting on introductory rates that double in 2027
Agents, avatars, media	Per second, per image, per generation, per tool call	Hard monthly ceilings, alert thresholds, and a documented unit rate for every modality in the order form	An open line the business discovers on the first invoice
Google Cloud commit	Committed spend over term with tiered discount; CUDs on resources	Commit sized from your own forecast, annual ramp, shortfall rollover, Marketplace drawdown, GPU and TPU CUD flexibility	Commit sized to Google's forecast with a fixed annual floor

Surface	What bills	Where the leverage is	The trap
Google SecOps	Committed GB per day with overage in arrears	Ingest cap sized to real telemetry, exempt sources, overage rate, bundling of SOAR, UEBA and intel at no incremental cost	Overage billed at list on log growth you never modeled
Mandiant and Threat Intelligence	Custom quote; retainer hours, subscriptions, per engagement	Benchmark against alternatives, unused retainer hours convertible, intel tier matched to sector	Buying the retainer in the week after an incident

The substance is the same across every surface: a metric the buyer does not yet control, a discount attached to a commitment, and a consumption line left open. What changes is whether the buyer arrives with a number of its own.

4. AGENTS, AVATARS AND MEDIA: THE OPEN LINE

The agent and media layer is where Google's pricing is least mature and the buyer's forecasting is weakest. It deserves its own treatment because it is the part of the deal most likely to be left open, and the part most likely to surprise finance.

- **Prebuilt agents inside Gemini Enterprise.** Deep research, data analysis, notebook and workflow agents draw on the seat's included quota and then on consumption. A single heavy user can consume the quota of ten light ones. Ask for the quota to be pooled across the estate and for the consumption rate to be fixed and capped, or the seat price is only the entry fee.
- **Custom agents on Vertex.** Anything built with the agent development kit or Agent Builder bills as tokens, tool calls, grounding requests and, where the agent runs continuously, compute. Each is a separate SKU. A production agent that runs several reasoning passes per task bills reasoning as output tokens, at the output rate, on every pass. Forecast on the production loop, not the demo.
- **Avatars and video presenters.** AI presenter and avatar features in Vids and the Gemini app, and video generation through Veo on Vertex, are metered per second of output at rates that make a minute of video cost more than a day of text for a typical user. Marketing, learning and internal communications teams adopt these fastest and forecast them least. Set the ceiling before the first campaign.
- **Image, music and speech generation.** Imagen, Lyria and text to speech bill per image, per generation or per character. They are small lines individually and unforecastable collectively. Treat them as one media ceiling with monthly alerts rather than as separate estimates.
- **Grounding and search.** Grounding agents on Google Search or on your own data bills per request on top of tokens. It is the line most often omitted from the proposal and most often present on the invoice.
- **Model change.** Every one of these surfaces will be repriced, deprecated or replaced within a three year term. The order form should carry the current rate, the tier, and a rule for what happens when the model behind it changes. Without that rule the price is whatever the successor model costs on the day of migration.

The practical position is simple: no agent or media line enters the agreement without a unit rate, a monthly ceiling, an alert threshold and a named owner. A line that meets those four tests can be bought with confidence. A line that does not should stay out of the commit until it does.

5. WHERE THE MONEY SITS

The illustration below is a mid sized enterprise, roughly 5,000 knowledge workers and an existing Google Cloud footprint, over a three year term at approximate list. The figures are for structure, not quotation.

Line	Google's opening structure	Approx. three year value	Buyer side structure	Approx. three year value
Gemini Enterprise seats	5,000 Standard seats, year one, full term	USD 5.4M	1,500 seats year one staged to 3,500 by year three, mix of Standard and Frontline, net 30 percent below list	USD 1.9M

Line	Google's opening structure	Approx. three year value	Buyer side structure	Approx. three year value
Seat consumption overage	Estimated, uncapped	USD 1.2M or open	Pooled quota across seats, capped at 15 percent of seat value, overage rate fixed	USD 0.5M
Vertex AI tokens	Forecast on introductory rates, standard rates from 2027	USD 3.6M	Price hold on current rates for term, volume tier from month one, batch and cached rates specified	USD 2.1M
Media and avatar generation	Open consumption	Unforecast	Monthly ceiling with approval gate, unit rates in order form	USD 0.4M
Google Cloud commit	USD 36M flat annual floor, three years	USD 36M	USD 27M ramped 7 / 9 / 11, shortfall rollover, 25 percent Marketplace drawdown, GPU flex CUDs	USD 27M
Google SecOps	600 GB per day cap, overage at list	USD 3.9M	400 GB per day with exempt sources, overage at committed rate, SOAR and UEBA included	USD 2.4M
Mandiant retainer and intel	Enterprise intel plus 400 hour retainer	USD 1.5M	Sector matched intel tier, 240 hours convertible to consulting, benchmarked	USD 0.9M
Total		**USD 51.6M plus open lines**		**USD 35.2M, all lines capped**

- **The commit is two thirds of the money.** Every seat and token discount in the proposal is small next to the difference between a floor sized to Google's growth model and a floor sized to your own forecast with a ramp and rollover.
- **Seats are the largest AI line and the easiest to stage.** Adoption data from a ninety day pilot is worth more than any discount argument. Buy the users who use it, with a documented right to add at the same net rate.
- **Tokens are the line that moves most after signature.** Model deprecation, successor pricing and the expiry of introductory rates change the unit cost without any change in your usage. The contract has to carry the rate, not the invoice.
- **Security is priced without a list.** Mandiant and Threat Intelligence quotes vary by a factor of three for similar scopes. Without an external benchmark the buyer has no way to know where the quote sits.
- **Open lines are the real risk.** Media generation, agent tool calls and SecOps overage are individually modest and collectively the difference between a budgeted deal and a surprise.

6. HOW BUYERS SHOULD NEGOTIATE

The response is a sequence. Its purpose is to establish your own numbers before Google's, unbundle the proposal into surfaces you can price, and keep every consumption line inside a ceiling you control.

- **1. Take the proposal apart.** Break the single number into its billing surfaces: Workspace edition uplift, agent seats, seat overage, tokens by model, media generation, developer seats, infrastructure commit, each security product. Price each on its own before allowing any of them back into a package.
- **2. Build your own forecast.** Pull Workspace usage reports, Gemini adoption telemetry, Vertex billing exports, SIEM ingest by source and Cloud billing by SKU. Forecast bottom up for three years. Know the number before Google tells you what it is.
- **3. Run a measured pilot before any seat commitment.** Ninety days, a defined population, an adoption threshold that decides who gets a seat. Bring the data to the table. Google discounts credible staging far more readily than volume promises.

- **4. Establish the alternative.** A live second AI provider through your other hyperscaler or a direct model vendor, with a real workload on it, is the only leverage Google respects on tokens and seats. For security, a benchmarked competing quote does the same work.
- **5. Size the commit from the forecast, then discount the forecast.** Commit to eighty percent of your bottom up number, ramped by year, with shortfall rollover, coterminous end dates and a documented mechanism for Marketplace drawdown. Never accept a flat floor on a growing AI estate.
- **6. Fix the unit rates for the term.** Token rates held past the 2026 introductory expiry, successor model pricing no worse than the current model at equivalent tier, seat rates held for additions, overage rates equal to committed rates. Put every rate in the order form.
- **7. Cap what you cannot forecast.** Monthly ceilings on media generation, agent consumption and seat overage, with alert thresholds and an approval gate. A ceiling you never hit costs nothing. An open line always costs something.
- **8. Price security on its own evidence.** Separate the SecOps ingest commitment from the cloud commit unless the drawdown is explicit and the rate is better. Benchmark Mandiant and intel quotes. Never buy a retainer during an incident; buy the engagement now and the retainer at renewal.
- **9. Negotiate the terms that decide year two.** True down rights at anniversary, model deprecation notice periods, data residency and zero data retention for prompts and outputs, the generative AI indemnity in writing, egress terms for training data, support tier pricing as a fixed amount rather than a percentage of the commit.
- **10. Sign on your calendar.** Google's quarter end is Google's leverage, not yours, unless you arrive with a signed alternative in hand. Set the timeline from your own renewal date and hold it.

7. CONTRACT TERMS THAT MATTER MOST

Term	What Google's paper says by default	What the buyer should hold out for
Token and model pricing	Rates per current price list; introductory rates expire on a stated date; successor models priced at Google's discretion	Rates fixed in the order form for the term; successor model at equivalent tier priced no higher; ninety days notice on deprecation with migration credit
Seat count and true down	Annual commitment, additions at then current price, no reduction until renewal	Additions at the same net rate; ten to twenty percent true down at each anniversary; edition downgrade permitted
Consumption overage	Billed at list in arrears; no cap	Overage at committed rate; monthly ceiling; alerting and approval gate; pooled quota across seats
Commit structure	Flat annual floor; shortfall invoiced; limited Marketplace drawdown	Ramped commit; shortfall rollover to the next year; full drawdown for first and third party Marketplace offers to the permitted cap
Data and IP	Standard data processing terms; indemnity language spread across service specific terms	No training on customer data; zero data retention for prompts and outputs where available; generative AI indemnity confirmed for the products in scope; data residency by region
Security products	Ingest cap fixed; overage at list; retainer hours expire	Exempt sources listed; overage at committed rate; unused retainer hours convertible; intel tier adjustable at anniversary
Support	Percentage of committed spend	Fixed annual amount, or percentage of consumed rather than committed spend

8. QUESTIONS TO PUT TO GOOGLE IN WRITING

Ask these before the proposal is finalized and keep the answers with the order form. Verbal answers from the account team do not survive the first invoice dispute.

- Which Gemini editions and which AI add ons are included in the Workspace price quoted, and what is the base price of the same edition without them?
- What is the included consumption quota per agent platform seat, is it pooled, what is the overage rate, and can a monthly ceiling be set at the tenant level?
- Which token rates in the proposal are introductory, when do they expire, and will the rates quoted be held for the full term in the order form?
- When a model in scope is deprecated, what is the notice period, what is the successor pricing rule, and is there a migration credit?
- What counts toward the commit: seats, tokens, media generation, Marketplace private offers from Google and from partners, and to what cap?
- What happens to an annual shortfall: invoiced, rolled forward, or applied against a later year, and on what terms?
- For SecOps, which sources are exempt from the ingest cap, what is the overage rate, and are SOAR, UEBA and threat intelligence included at the quoted tier?
- Which data processing terms, data residency options, retention settings and indemnities apply to each product in scope, and where are they written down?
- What is the support charge as a fixed amount, and does it move with committed or consumed spend?

9. MISTAKES THAT COST THE MOST

- Signing agent platform seats for the whole workforce on the strength of a demo, before adoption data exists.
- Building the three year token budget on introductory rates and discovering the standard rates on the first 2027 invoice.
- Accepting a flat commit floor sized to Google's growth model because the year one discount looked generous.
- Leaving media generation, avatar and agent consumption uncapped because nobody could forecast them.
- Buying the Mandiant retainer and the SecOps migration in the same week as the incident response engagement.
- Letting the security team and the AI team negotiate separately with the same Google account team, which then sizes one commit from both.
- Treating the Google proposal as the only proposal. Without a second provider running a real workload, every discount is a favor.

10. HOW REDRESS HELPS

Redress Compliance runs the exercise from the buyer's side: the proposal unbundled into its billing surfaces, the bottom up forecast from your own telemetry, the pilot design and adoption thresholds, the commit and ramp model, the benchmark on seats, tokens and security quotes, the contract terms that decide year two, and the negotiation itself on your calendar. Our fees are fixed, and every dollar removed from Google's number belongs to the client.

REDRESS COMPLIANCE LLC

1314 E Las Olas Blvd, Fort Lauderdale, FL 33301 | redresscompliance.com | +1 (239) 402-7397

100 percent buyer side. Zero vendor affiliations. No reseller agreements. No referral fees.