

ENGINEERING WHITE PAPER

How Dovely Chooses a Model

Model routing at Edenly: principles, evidence, and policy

Edenly Engineering Team

Tech of Eden PBC

Version 1.0 · September 2026

Every figure in this document is copied from a dated, public source listed in Appendix B. Nothing is estimated. Where a figure does not exist, we say so.

Executive summary

Dovely answers with the help of large language models made by several companies. This document explains, in plain terms and with sources, how we decide which model answers a given question.

The headlines — the two switches worth remembering.

- **Crisis and grief go to Claude.** On the one emotional-intelligence benchmark every one of our candidates took, Claude's latest model wins by a distance — so the moment the Artificial Conscience flags a crisis or grief conversation, **Claude's latest model** writes the answer (Section 7.2).
- **Code goes to ChatGPT.** For technical questions the writer is **ChatGPT's latest everyday model**, which leads the default on the coding benchmark at no extra cost per answer (Section 6.3).
- **Everything else runs on the everyday tier**, where the independent scores put the field level — and every answer, whichever company's model writes it, passes the same Artificial Conscience review before it reaches you. We pull the best from the different models; no single vendor is the story.

The short version:

1. **Every answer goes through the Artificial Conscience first, whatever model writes it.** The moral and pastoral guardrails are ours; they do not depend on the model vendor.
2. **You never pick a model.** Dovely reads your question, classifies it, and routes it. The everyday tier handles the vast majority of traffic.
3. **Models are grouped into three capability tiers** — Everyday, Plus and Frontier — cut on one independent composite score (Section 5). Within a tier the models are interchangeable on general quality, so we choose on cost and speed. Across tiers we move only on evidence.
4. **We route on evidence, not on brand.** An activity is routed to a higher tier only when public benchmarks show a meaningful, measurable advantage for that activity *and* the activity's share of traffic justifies the cost.
5. **One activity is routed to the Frontier tier: crisis and grief.** There the writing model is Claude's latest, which leads the field on the one independent emotional-intelligence benchmark that covers our candidates by about 104 Elo over Claude Sonnet 5 and 184 Elo

over the everyday field, with non-overlapping confidence intervals. Everything else stays on the everyday tier.

6. **Savings are yours.** Using the most economical model wherever quality is equivalent is what lets each plan carry more usage for the same price. The everyday chain is built of the cheapest models that hold the tier's quality (Section 5).

This document is maintained by the engineering team and re-issued when the routing table changes.

1. Why we are publishing this

People rightly ask two questions of any AI product: *can I trust what it says*, and *is it any good at the job*. The Artificial Conscience answers the first. This document answers the second, honestly: which model does the work, why, and what evidence that decision rests on.

We would rather show the table than ask for faith in a slogan.

2. Principles

The Conscience is model-independent. The Artificial Conscience — the policy layer that reads every message, applies the relevant moral and pastoral policies, reviews every draft, and rewrites when a draft fails a check — runs around the model, not inside it. Switching models changes the writer, never the rules the writer must follow.

No model picker. A picker would move the burden of expertise onto you and would make the product's behaviour depend on a setting most people never touch. Routing is the product's job.

Evidence over brand. Model vendors publish results; independent leaderboards publish more. We use the comparable, dated, sourced figures — and we say when a figure is vendor-reported rather than independently measured.

Meaningful differences only. Benchmarks carry noise. A one-point gap on a hundred-question test is not a reason to change anything. We route only when an activity shows a difference larger than the benchmark's own uncertainty *and* large enough to matter to the person asking.

Cost is stewardship. We pay for every token the models produce. Using the most economical model wherever quality is equivalent is what keeps the plans priced as they are and what lets us give each plan more room.

3. How Dovely processes a message

Every message follows the same path regardless of which model eventually writes the answer:

1. **Classification (the AC gate).** Small, fast classification questions run on the incoming message in parallel: whether there is any risk of harm and how severe, which policy domains apply, what state the person seems to be in, whether Scripture or trusted Christian references should be consulted, and other signals the policies need. The stage is inexpensive, runs on every message, and produces a set of labels.
2. **Reflection.** A short reflective pass considers what a wise response would weigh — the heart of the question, not only its surface — and may add policy labels the first pass missed.
3. **Policy composition.** The labels select the applicable Artificial Conscience policies (and, for members of a church that has configured its positions, that church's stances). These are composed into the instructions the writing model receives.
4. **Generation.** A model writes the answer under those instructions. *This is the only step where routing chooses between models.*
5. **Output review.** Each applicable policy grades the draft against its own check questions. A failed check sends the draft back with the specific failures; the writer revises; the review runs again.
6. **Delivery.** Only a draft that passed review reaches you.

Routing therefore has a narrow, well-defined job: at step 4, pick the writing model for this activity. Steps 1–3 and 5–6 are identical for every model.

Steps 1, 2 and 5 are themselves model calls — several of them per user message. They are structured, machine-read work, not prose the person ever sees, and they are routed on their own terms (Section 7.2, first row).

4. What routing means here

Routing is a table: **activity** → **writing model (with an ordered fallback list)**. It is not personal (it does not depend on who you are), it is not commercial (no vendor pays for placement), and it is not a preference setting.

The table has a **default row** that handles every question not matched by a more specific row. The default is the model that, across the activities people actually bring to Dovely, sits in the same capability tier as the field's best while costing a fraction of it. Today that is DeepSeek V4 Flash.

Every other row must earn its place with the evidence in Section 6.

The activities are detected from the labels the Artificial Conscience gate already produces on every message, plus a small number of *technical* signals added to that same pass — no extra round-trip, no extra cost. Activities are mutually exclusive for routing purposes; when several match, the one with the strongest benchmark evidence for a higher tier wins, otherwise the default applies. A misclassified message is not a failure mode for safety — the Conscience applies either way — only, at worst, a slightly less capable writer for that one answer.

5. Model tiers

5.1 How the tiers are cut

We group the nine models we can compare into three **capability tiers** using one independent composite: the Artificial Analysis Intelligence Index v4.1.1 (S25, S35). It is a single number produced by a third party across a fixed battery of evaluations, run on every model in the same harness, which is why we use it for the grouping and use the class-specific benchmarks in Section 6 for everything else.

The cuts fall where the gaps in the published scores fall: 52 · 55–56 · 61–63.

Tier	Model	AA index	Arena text overall	List price in / out (\$/1M)	AA blended (\$/1M)	Output tok/s	First answer (s)
1 · Everyday	DeepSeek V4 Flash (default)	52	1436 ±4 (89)	0.44 / 1.32 peak; 0.22 / 0.66 off-peak	0.23	107.9	1.50
1 · Everyday	GPT-5.6 Luna	52	1452 ±5 (63)	0.20 / 1.20	0.17	129.8	168.55
1 · Everyday	Gemini 3.5 Flash	52	1479 ±4 (21)	1.50 / 9.00	1.31	196.6	16.26
2 · Plus	Gemini 3.7 Flash	56	1490 ±8 (9)	0.75 / 3.75; 1.50 / 7.50 from 2027	0.58	289.5	9.47
2 · Plus	Claude Sonnet 5	55	1461 ±5 (47)	2.00 / 10.00	1.54	73.3	182.62
3 · Frontier	Claude Opus 5	63	1492 ±5 (7)	5.00 / 25.00	3.85	53.7	47.93
3 · Frontier	Claude Fable 5	62	1507 ±5 (1)	10.00 / 50.00	7.70	66.9	128.51
3 · Frontier	GPT-5.6 Sol	61	1482 ±5 (17)	4.00 / 20.00	3.08	77.1	115.66
3 · Frontier	Grok 4.6	61	1461 ±10 (48)	2.00 / 6.00	1.35	54.4	44.74

Sources: AA Intelligence Index v4.1.1, AA blended price, output speed and time to first answer token — S25 (the original six) and S35 (Sol, Fable 5, Gemini 3.7 Flash), retrieved 1 September 2026. Arena text overall — S24, snapshot 27 August 2026. List prices — S4 (DeepSeek), S7

(OpenAI), S13 (Google), S17 (Anthropic), S23 (xAI), retrieved 1 September 2026. Arena cells carry the score $\pm$ 95% confidence interval with the board rank in brackets. Gemini 3.7 Flash's list price is introductory through 31 December 2026; the second pair applies from 1 January 2027.

One note on reading the table. **First answer** is Artificial Analysis's time to first answer token, measured at high or maximum reasoning effort and including thinking time; it is not chat-default latency, and the long values for Luna, Sonnet 5, Sol and Fable 5 reflect that measurement, not what a person waits for in a conversation.

5.2 The rule inside a tier

Within a tier, the models are interchangeable on general quality. They sit within a few points of each other on the composite and their conversational scores overlap or nearly overlap. When that is true, the honest statement is not "model X is best" but "these are the same tier for this work", and the pick is made on cost and speed.

The exception is a class-specific benchmark that separates them. Inside Tier 3, EQ-Bench 4 does exactly that: Claude Fable 5 leads GPT-5.6 Sol by 86.9 Elo with non-overlapping intervals (Section 6.7), so for the one activity where emotional handling is the whole job, the tier's cheapest member is not automatically the pick.

One consequence is worth stating plainly. **Gemini 3.7 Flash is a Tier 2 model priced like Tier 1**, and it dominates the Tier 1 Gemini 3.5 Flash beside it on every measure: index 56 vs 52, list price 0.75 / 3.75 vs 1.50 / 9.00, blended price 0.58 vs 1.31, 289.5 vs 196.6 tokens per second, first answer in 9.47 s vs 16.26 s. When its introductory pricing ends on 31 December 2026 its input price becomes equal to Gemini 3.5 Flash's and its output price stays 17% below, so the dominance survives the increase.

5.3 Model tiers are not plan tiers

Edenly's paid plans are also called tiers — \$5 plugin-only, \$10 and \$20 — and they are a different thing entirely. **Plans meter usage in credits. They never change which model answers.**

A \$5 subscriber asking about grief and a \$20 subscriber asking about grief get the same writing model, because the activity is what selects the tier, not the wallet. What the plan buys is how much you can ask, not who answers.

This matters for the cost arithmetic in Section 9. Because a Frontier-tier answer costs us roughly 35× an everyday answer, routing an activity up is a decision about the plans everyone pays for — not a premium feature we can sell to the people who need it most.

6. Evidence

All figures were retrieved on 1 September 2026. **V** = reported by the model's own vendor; **X** = reported by another vendor about this model; **I** = independent leaderboard or evaluator; **T** = third-party rendering of a vendor document we could not parse directly. Source IDs are resolved in Appendix B.

The nine models compared are DeepSeek V4 Flash (served build 0731), GPT-5.6 Luna, GPT-5.6 Sol, Gemini 3.5 Flash, Gemini 3.7 Flash, Claude Sonnet 5, Claude Opus 5, Claude Fable 5 and Grok 4.6.

6.1 General knowledge & reasoning

Model	Metric	Value	Variant	Type	Source
DeepSeek V4 Flash 0731	Humanity's Last Exam, no tools / with tools	37.8 / 51.5	shipped build	V	S4
Gemini 3.5 Flash	Humanity's Last Exam	40.2%	full set, text + multimodal	V	S11
Claude Sonnet 5	Humanity's Last Exam, no tools / with tools	43.2% / 57.4%	—	V	S15
Claude Opus 5	Humanity's Last Exam	54.9%	adaptive reasoning, max effort	I	S26
Claude Fable 5	Humanity's Last Exam	55.5%	adaptive reasoning, max effort, Opus 4.8 fallback	I	S26
Gemini 3.7 Flash	HLE-Verified	53.6%	Google's own variant, on its own card	V	S14
Claude Sonnet 5	HLE-Verified	31.0%	same table, as run by Google	X	S14
GPT-5.6 Sol · Gemini 3.7 Flash	Humanity's Last Exam (standard)	not published; Google publishes only the "HLE-Verified" variant	—	gap	S5, S14

The Artificial Analysis Intelligence Index v4.1.1 score for each of the nine is in the tier table (Section 5.1), each model run at its high or maximum reasoning setting (S25, S35). Two classes support no comparison and are recorded here rather than tabulated. MMLU-Pro is published only by DeepSeek (83.0 / 86.4 / 86.2 EM at its three reasoning settings, S2). GPQA Diamond is published for six models across three harnesses, which is why we do not rank on it: DeepSeek V4

Flash 88.1 at max effort (S2); Luna 92.3, Sol 94.6 and Fable 5 92.6 as run by OpenAI (S5); Sol 94.1, Gemini 3.7 Flash 94.5 and Grok 4.6 94.9 as run by Artificial Analysis (S26). Humanity's Last Exam is not published for Luna, Gemini 3.5 Flash or Grok 4.6.

What the numbers show. On the one independent composite that covers all nine models the tiers are clean: three models at 52, two at 55–56, four at 61–63. Fable 5 (62) and Sol (61) are within a point of each other and just under Opus 5 (63); Gemini 3.7 Flash's 56 is four points above Gemini 3.5 Flash. On Humanity's Last Exam without tools the shipped DeepSeek Flash (37.8) is 5.4 points under Sonnet 5, 17.1 under Opus 5 and 17.7 under Fable 5.

6.2 Mathematical / quantitative reasoning

Model	Metric	Value	Variant	Type	Source
DeepSeek V4 Flash	HMMT February 2026 (Pass@1)	40.8 / 91.9 / 94.8	Non-Think / High / Max	V	S2
DeepSeek V4 Flash	IMOAnswerBench (Pass@1)	41.9 / 85.1 / 88.4	Non-Think / High / Max	V	S2
GPT-5.6 Sol	FrontierMath Tiers 1–3 (v2)	89.1% $\pm$ 1.8	max	I	S27
Claude Fable 5	FrontierMath Tiers 1–3 (v2)	87.0% $\pm$ 2.0	max	I	S27
Claude Opus 5	FrontierMath Tiers 1–3 (v2)	85.6% $\pm$ 2.1	max	I	S27
GPT-5.6 Luna	FrontierMath Tiers 1–3 (v2)	82.1% $\pm$ 2.3	max	I	S27
Grok 4.6	FrontierMath Tiers 1–3 (v2)	66.0% $\pm$ 2.8	xhigh	I	S27
Claude Sonnet 5	FrontierMath Tiers 1–3 (v2)	65.6% $\pm$ 2.8	max	I	S27
Gemini 3.7 Flash · Gemini 3.5 Flash · DeepSeek V4 Flash	FrontierMath	no row published or retrievable from any source checked	—	gap	S27, S14, S41

Not published: AIME 2025/2026 and MATH-500 for any of the nine from their vendors.

What the numbers show. The only independent, like-for-like math row for most of the field is Epoch's FrontierMath: Sol leads at 89.1%, then Fable 5 87.0%, Opus 5 85.6%, Luna 82.1%, with Grok 4.6 and Sonnet 5 twenty points back and statistically tied with each other. On the hardest

tier the order between the top two reverses on OpenAI's own card (Fable 5 87.8% against Sol 83%, S5). DeepSeek publishes different competition benchmarks (HMMT, IMOAnswerBench) with no overlap, and neither Gemini has a row at all, so **no math figure is directly comparable between the default model and the others** — the routing rule cannot be applied to this activity from public data alone.

6.3 Code & technical implementation

Model	Metric	Value	Variant	Type	Source
DeepSeek V4 Flash	SWE-bench Pro (Resolved)	49.1 / 52.3 / 52.6	Non-Think / High / Max	V	S2
GPT-5.6 Luna	SWE-bench Pro · Terminal-Bench 2.1 · DeepSWE v1.1	62.7% · 84.7% · 67.2%	—	V	S5
GPT-5.6 Sol	SWE-bench Pro · Terminal-Bench 2.1 · DeepSWE v1.1	64.6% · 88.8% · 72.7%	—	V	S5
GPT-5.6 Sol	SWE-bench Verified	96.20%	Vals AI bash-only harness, rank 1	I	S36
Gemini 3.5 Flash	SWE-bench Pro (Public) · Terminal-Bench 2.1	55.1% · 76.2%	single attempt	V	S11
Gemini 3.7 Flash	Terminal-Bench 2.1 · FrontierCode 1.1 · DeepSWE v1.1 · Code Arena web dev (Elo)	85.8% · 43.6% · 65.3% · 1588	—	V	S14, S41
Claude Sonnet 5	SWE-bench Pro · Terminal-Bench 2.1	63.2% · 80.4%	—	V	S15
Claude Opus 5	SWE-bench Verified	97.00%	Vals AI bash-only harness, rank 1	I	S29
Claude Opus 5	SWE-bench Pro	79.2%	as summarised from the system card	T	S19
Claude Fable 5	SWE-bench Pro	80%	as run by OpenAI	X	S5
Claude Fable 5	SWE-bench Verified	95.00%	Vals AI bash-only harness, rank 1	I	S36
Claude Fable 5	Terminal-Bench 2.1	80.52% · 88.0%	Vals AI re-run · as run by DeepSeek	I, X	S36, S4
Grok 4.6	DeepSWE v1.1 (Pass@1)	65.9% / 67.0%	high / xhigh	V	S22

Model	Metric	Value	Variant	Type	Source
Gemini 3.7 Flash	SWE-bench Verified · SWE-bench Pro	not published, and no independent row	—	gap	S14, S41, S36

Not published: SWE-bench Verified and LiveCodeBench for Luna, Gemini 3.5 Flash, Sonnet 5 and Grok 4.6 from their vendors; the official SWE-bench and LiveCodeBench leaderboards carry no 2026 entries for these models.

What the numbers show. On SWE-bench Pro the ordering is Fable 5 80.0 and Opus 5 79.2 at the top, then Sol 64.6, Sonnet 5 63.2 and Luna 62.7 within two points of each other, then Gemini 3.5 Flash 55.1 and DeepSeek V4 Flash at max effort 52.6 — each figure vendor-reported or vendor-run, harnesses may differ. Luna's lead over the default is 10.1 points; the Frontier tier's lead over Luna is a further 16–17. Terminal-Bench is not a routing signal: Fable 5's own figure moves 7.5 points depending on who ran it (80.52 Vals AI, 88.0 DeepSeek). On the one independent same-harness comparison (SWE-bench Verified, Vals AI) the three Frontier models sit inside a two-point band, from Opus 5 at 97.00 to Fable 5 at 95.00.

6.4 Instruction following

No IFEval or IFBench figure is published for any of the nine models by their vendors, and none appears in the independent sources that could be read. **Gap — no routing decision can rest on this class.**

6.5 Long-context work

Eight of the nine models take an input window of a million tokens or more: DeepSeek V4 Flash 1M (S1), both GPT-5.6 models 1.05M with 128K output (S6, S37), both Gemini Flash models 1M input with 64K output (S11, S14), and Claude Sonnet 5, Opus 5 and Fable 5 1M with 128K output, the 1M window billed at the standard rate (S17, S40). Grok 4.6 takes 500K (S23).

Model	Metric	Value	Variant	Type	Source
DeepSeek V4 Flash	MRCR at 1M (MMR)	37.5 / 76.9 / 78.7	Non-Think / High / Max	V	S2
GPT-5.6 Luna	OpenAI MRCR v2, 8-needle, 256K–512K · 512K–1M	41.3% · 41.3%	—	V	S5
GPT-5.6 Sol	OpenAI MRCR v2, 8-needle, 256K–512K · 512K–1M	91.5% · 73.8%	—	V	S5
Gemini 3.5 Flash	GDM-MRCR v2, 8-needle, 128k average · 1M pointwise	77.3% · 26.6%	—	V	S11

Model	Metric	Value	Variant	Type	Source
Gemini 3.7 Flash	GDM-MRCR v2, 8-needle, 128k average	97.0%	—	V	S14
Claude Sonnet 5	GDM-MRCR v2, 8-needle, 128k average	81.5%	as run by Google	X	S14
Claude Fable 5	MRCR or any needle benchmark	not published in any source checked	—	gap	S38, S40

What the numbers show. Every MRCR figure uses a different window or protocol (DeepSeek at 1M MMR, OpenAI v2 at 256K–1M, Google v2 at 128k), so **none are comparable across vendors**. Within Google's own 128k protocol three models are directly comparable: Gemini 3.7 Flash 97.0%, Sonnet 5 81.5%, Gemini 3.5 Flash 77.3%. The default model has only a 1M-window MMR figure (78.7 at max effort) under a protocol nobody else runs, so **no like-for-like comparison against the default exists**.

6.6 Multilingual

DeepSeek publishes base-model MMMLU (88.8, 5-shot) and MGSM (85.7); no post-trained multilingual knowledge benchmark is published for any of the nine chat models, and none of the three frontier candidates appears in Artificial Analysis's Global-MMLU-Lite rows (S26). **Gap — no routing decision can rest on this class.**

6.7 Conversational quality & writing

Arena (formerly LMArena) category boards, snapshot 27 August 2026 (S24, independent). Each cell carries the score $\pm$ 95% confidence interval with the board rank in brackets. Text overall scores and ranks for all nine models are in the tier table (Section 5.1); the order below follows them.

Model	Creative Writing	Longer Query	Multi-Turn
claude-fable-5	1505 $\pm$ 9 (1)	1521 $\pm$ 7 (2)	1515 $\pm$ 10 (4)
claude-opus-5-high	1471 $\pm$ 9 (10)	1506 $\pm$ 7 (6)	1483 $\pm$ 13 (27, max variant)
gemini-3.7-flash-high	1494 $\pm$ 18 (3, preliminary)	1499 $\pm$ 12 (13, preliminary)	1498 $\pm$ 21 (9)
gpt-5.6-sol-xhigh	1477 $\pm$ 10 (9)	1495 $\pm$ 7 (14)	1490 $\pm$ 11 (18)
gemini-3.5-flash-high	1465 $\pm$ 9 (15)	1482 $\pm$ 6 (34)	1485 $\pm$ 9 (24)
claude-sonnet-5-high	1437 $\pm$ 9 (51)	1482 $\pm$ 6 (35)	1471 $\pm$ 9 (48)
grok-4.6-high	1456 $\pm$ 24 (25)	1479 $\pm$ 15 (36)	1445 $\pm$ 25 (86)

Model	Creative Writing	Longer Query	Multi-Turn
gpt-5.6-luna-xhigh	1409 ±10 (79)	1452 ±7 (73)	1456 ±10 (68)
deepseek-v4-flash	not on the board	not on the board	1452 ±7 (72)

Emotional-intelligence evaluation, EQ-Bench 4 official results (S30, independent, Elo with 95% CI):

Model	Elo	CI
claude-fable-5	1349.50	1335.05–1365.25
gpt-5.6-sol	1262.59	1252.49–1273.75
claude-sonnet-5	1245.86	1236.38–1256.63
deepseek-v4-pro	1176.93	1166.48–1186.48
gpt-5.6-luna	1165.77	1151.19–1177.96
gemini-3.5-flash	1097.50	1084.80–1106.96
deepseek-v4-flash · claude-opus-5 · grok-4.6 · gemini-3.7-flash	not in the results file	—

What the numbers show. On overall conversational preference the nine span 71 Elo, with the default model (DeepSeek V4 Flash, 1436) at the bottom and Claude Fable 5 (1507) at the top of the board. Fable 5 is top-four on every board Arena publishes for it: rank 1 overall, rank 1 Creative Writing, rank 2 Longer Query (1521), rank 4 Multi-Turn (1515). Gemini 3.7 Flash ranks 3rd–13th across the four boards, its Creative Writing and Longer Query placements marked preliminary by Arena.

On emotional intelligence the picture sharpens. Claude Fable 5 (1349.50) leads Sonnet 5 (1245.86) by 103.6 Elo, Sol (1262.59) by 86.9 and Luna (1165.77) — the highest-scoring Tier 1 model present — by 183.7, in every case with non-overlapping confidence intervals. Sol and Sonnet 5 overlap and are treated as tied. The default model's Flash build is not in the EQ-Bench file (its Pro sibling overlaps Luna within CI), and Gemini 3.7 Flash is absent from it entirely, so we hold no emotional-intelligence figure for either.

6.8 Cost and speed

Official list prices per one million tokens, and the independent output speed, time to first answer token and blended price, are in the tier table (Section 5.1): vendor pages S4, S7, S13, S17 and S23 for the prices, Artificial Analysis S25 and S35 for the measured figures, all retrieved 1 September 2026. The terms that do not fit in that table: DeepSeek's peak window is 01:00–04:00 and 06:00–10:00 UTC, with the off-peak prices otherwise (S3); Sol's cached input is \$0.40 and

Fable 5's is \$1.00, with batch runs at half list for Fable 5 (S7, S17); Gemini's prices are the paid tier with thinking included, and Gemini 3.7 Flash's batch and flex tiers are half list (S13); Claude Sonnet 5's introductory price is permanent (S17).

What the numbers show. Per output token, Luna (\$1.20) is below DeepSeek at peak (\$1.32) and is the cheapest model of the nine on input; Gemini 3.7 Flash output costs 2.8× DeepSeek peak, Grok 4.6 4.5×, Sonnet 5 7.6×, Gemini 3.5 Flash 6.8×, Sol 15.2×, Opus 5 18.9× and Fable 5 37.9×. Gemini 3.7 Flash is both the fastest generator and the cheapest on the independent blended metric of all nine models; DeepSeek has by far the lowest first-answer latency at max effort (1.50 s, the next best being 9.47 s). Claude Fable 5 is the most expensive model in the set on every price measure — list, blended, and per answer.

6.9 How to read these tables

- **Comparable variants only.** Where a benchmark has several protocols we name the variant; figures with different protocols are shown but never compared as if equal.
- **Vendor-reported vs independent.** Vendor figures are useful and usually accurate; independent figures are preferred whenever both exist. Where sources disagree on the same benchmark for the same model (Fable 5 on Terminal-Bench 2.1), we show the disagreement rather than pick a number.
- **Uncertainty.** Where a source publishes confidence intervals (Arena, Epoch, EQ-Bench), two models whose intervals overlap are treated as tied.
- **Gaps are findings.** Where an activity has no comparable figure, we say so and do not route on it.

7. Routing policy

7.1 The rule

An activity is routed to a tier above the default only when **all** of the following hold for a candidate model, on the benchmark(s) representative of that activity:

1. the candidate leads the default by a margin larger than the benchmark's published uncertainty (or, where none is published, by at least five points on a 100-point scale);
2. the advantage appears on more than one relevant benchmark or in more than one independent source;
3. the activity is detectable reliably enough that the routing decision is right far more often than wrong (we measure this on a held-out set of real, anonymised classifications);
4. **the activity's share of traffic justifies the tier's cost.** A Frontier answer costs us roughly 35× an everyday answer (Section 9). An advantage that is real but applies to a large share of traffic is not automatically worth buying; an advantage on a small, high-stakes share is.

If any condition fails, the activity stays on the default tier.

7.2 The table

Activity	Tier	Writing model → fallback chain	Detection signal	Why this tier
Internal AC gate steps — input labels, checks, reflection, rewrite verdicts	1 · Everyday, cheapest available	DeepSeek V4 Flash → GPT-5.6 Luna → Gemini 3.7 Flash	Not user-facing; runs on every message	Several calls per user message. These produce structured labels and verdicts, verified against our own fixture corpus, not prose anyone reads. Benchmark deltas between models do not transfer to this work; cost dominates.
Everyday faith, Bible and life questions	1 · Everyday	DeepSeek V4 Flash → GPT-5.6 Luna → Gemini 3.7 Flash — <i>the default chain</i>	Default — everything not matched below	Index parity across the tier (52 = 52 = 52, \$6.1); the bulk of all traffic; lowest cost and lowest first-answer latency (\$6.8).
Pastoral, relationship and decision guidance — no crisis signal	1 · Everyday	Default chain	AC policy-domain labels without an elevated harm-severity or crisis label	A cost decision, stated as one. See the paragraph below the table.
Crisis, grief, elevated harm severity	3 · Frontier	Claude Fable 5 → GPT-5.6 Sol → Claude Sonnet 5	The AC gate's harm-severity and crisis/grief labels only	EQ-Bench 4: Fable 5 1349.50 (CI 1335.05–1365.25) leads Sonnet 5 by 103.6 Elo, Sol by 86.9 and the everyday field by 183.7, intervals disjoint (\$6.7). Arena: rank 1 text overall (1507 ±5), Longer Query 1521, Multi-Turn 1515. A small share of traffic and exactly where quality matters most.

Activity	Tier	Writing model → fallback chain	Detection signal	Why this tier
Code & technical implementation — provisional	1 · Everyday	GPT-5.6 Luna → DeepSeek V4 Flash	Code fences, language names, stack traces, implement/refactor/debug intent	Luna leads the default by 10.1 on SWE-bench Pro (62.7 vs 52.6, each vendor-reported, §6.3) at no more cost per output token. The Frontier tier leads further (Fable 5 80.0, Opus 5 79.2) but code is a negligible share for a faith companion and never high-stakes for the person; 20–40× cost is not justified.
Long-document work	1 · Everyday	Default chain	Input length past a threshold plus a document-operation intent	No like-for-like figure against the default exists (§6.5). Gemini 3.7 Flash is the first candidate to test: 97.0% on GDM-MRCR 128k against Sonnet 5's 81.5% on the same Google-run table, while the default has only a 1M-window figure under a protocol nobody else runs.
Creative writing — prayers, reflections, devotionals	1 · Everyday	Default chain	Explicit creative-form request	Arena Creative Writing has Fable 5 at 1505 and Gemini 3.7 Flash at 1494 (preliminary), but the default is not on that board at all — no delta is measurable (§6.7).
Multilingual	1 · Everyday	Default chain	Language detection on the message	No post-trained multilingual knowledge benchmark is published for any of the nine (§6.6).
Mathematical / quantitative	1 · Everyday	Default chain	Formal quantitative structure the person needs solved, not merely mentioned	GPQA Diamond is science QA, not maths; FrontierMath has no row for the default or for either Gemini (§6.2). No like-for-like comparison exists.

On the pastoral row, honestly. This is the activity where the gate's "Jesus would pause" judgement fires most, and it is the activity a reader would most expect to find in the Frontier tier. It is not there, and the reason is cost, not a claim that the models are equal. EQ-Bench 4 shows the higher tiers are measurably stronger at emotionally loaded conversation — Sonnet 5 leads Luna by 80.1 Elo — and we have no reason to think that advantage stops at the crisis boundary. What stops us is arithmetic: pastoral guidance without a crisis signal is a large share of everything people bring to Dovely, and moving that share up a tier multiplies the cost of the bulk of the product by 7× (Sonnet 5) to 35× (Fable 5) per answer. Two things make that trade acceptable today. First, the Artificial Conscience's output review enforces the pastoral requirements on every answer whichever model wrote it, so the floor does not move with the writer. Second, this is a measurable question, not a matter of taste: we are instrumenting the output-review rewrite rate for this activity, and if the everyday tier needs materially more rewrites here than elsewhere, the row moves up (Section 11).

On the crisis row, two notes. The detection signals are the harm-severity and crisis/grief labels the Artificial Conscience gate already produces on every message — routing adds no classification cost. The gate's "Jesus would pause" judgement is deliberately **not** a routing signal: it fires on a wide range of ordinary questions (doctrine, relationships, money, decisions) and would send everyday traffic to the most expensive tier. Only crisis-grade signals move a conversation. Second, the chain deliberately crosses tiers — Sol is Tier 3, Sonnet 5 is Tier 2 — because at the third link an answer written by a slightly weaker model under the same Conscience is better than no answer, and Frontier models are no more available than everyday ones (Section 10).

7.3 Expected effect

The routed activities are a small minority of traffic and the everyday tier writes the overwhelming majority of Dovely's answers. That is by design: routing exists to keep the ceiling high where it matters, not to spread traffic across vendors. The actual shares are not yet measured (Section 9). is five percent or less — and the everyday tier writes the overwhelming majority of Dovely's answers. That is by design: routing exists to keep the ceiling high where it matters, not to spread traffic across vendors. The actual shares are not yet measured (Section 9).

8. Same tier, our choice

Inside a tier the models are tied on the independent composite and separated by a few Elo on general preference (Section 5.2), so the choice between them is ours to make on cost, speed, reliability and data handling. That is what makes this document durable.

It is also why the default can change without the routing logic changing. If the Google infrastructure credits we have applied for land, the default row moves from DeepSeek V4 Flash to Gemini 3.7 Flash — a **higher** tier at a lower cost to us, requiring no change to any policy in Section 7. The everyday activities stay on the default row, the crisis row stays on the Frontier tier,

and the rows that exist because of a measured advantage are unaffected by which model holds the default. The same applies in the other direction: if a provider's price rises, its reliability drops or its terms change, we move within the tier and nothing about what Dovely does changes.

9. Cost and speed stewardship

Cost per answer is derived arithmetic from the list prices in Section 6.8, at an **assumed message size of 400 input tokens and 600 output tokens**. That size is a stated assumption, not a measurement; it is used only to put the tiers on one scale. Figures below are USD per 1,000 answers, before any batch, cached-input or off-peak discount.

Tier	Model	\$ per 1,000 answers	Multiple of the default
1 · Everyday	DeepSeek V4 Flash, off-peak	0.48	0.5×
1 · Everyday	GPT-5.6 Luna	0.80	0.83×
1 · Everyday	DeepSeek V4 Flash, peak (the default)	0.97	1.0×
1 · Everyday	Gemini 3.5 Flash	6.00	6.2×
2 · Plus	Gemini 3.7 Flash (introductory)	2.55	2.6×
2 · Plus	Gemini 3.7 Flash (from 1 Jan 2027)	5.10	5.3×
2 · Plus	Claude Sonnet 5	6.80	7.0×
3 · Frontier	Grok 4.6	4.40	4.5×
3 · Frontier	GPT-5.6 Sol	13.60	14.0×
3 · Frontier	Claude Opus 5	17.00	17.6×
3 · Frontier	Claude Fable 5	34.00	35.1×

Three consequences follow:

- **The gap between tiers is a multiple, not a margin.** Routing every question to the Frontier tier would cost us about thirty-five times what the everyday tier costs, and would make the plans either far more expensive or far more restricted. Routing only the crisis activity keeps the blended cost within a few percent of the default.
- **We will not publish a blended figure until we have measured the shares.** A claim like "routing costs us 3% more" would require knowing what fraction of traffic is crisis-labelled, and we do not yet know it. Section 11 commits to measuring it; this section will carry the number when it exists.

- **Speed matters to the experience.** Output speed and first-answer latency are part of the evidence we weigh, particularly for the conversational activities. The default model begins answering in 1.50 seconds at maximum effort, six times faster than the next model measured.

10. Safety, privacy and the Conscience

- **Policy independence.** The Artificial Conscience's policies, church overlays and output review are applied identically for every writing model. A model is never allowed to "opt out" of a policy check.
- **Data handling.** We send each provider only what it needs to write the answer, under that provider's API terms for business customers. Provider data-retention terms are referenced in the appendix and re-checked when the routing table changes.
- **Failure handling.** Every route carries an ordered fallback chain; if a model is unavailable or returns an unusable answer, the next in the chain writes under the same instructions and the same review. Models are occasionally unavailable — Claude Fable 5 was unavailable to all users between 12 June and 1 July 2026 (S38, S39) — which is why no route depends on a single model.
- **An honest gap in the safety evidence.** No independent or vendor-published study compares any of our candidate models against another on harm handling, sycophancy or emotional-support quality. Vendors publish their own safety suites — OpenAI's mental-health and emotional-reliance evaluations (S8, S9), Google's safety deltas (S11, S14), xAI's sycophancy and self-harm compliance rates (S22) — with different metrics and different polarity, which cannot be compared across vendors. EQ-Bench 4 is the only cross-vendor, like-for-like signal we have, and it measures emotional intelligence in conversation, not crisis safety. The crisis route therefore rests on the best proxy available plus our own output review — not on a benchmark that measures the thing itself, because no such benchmark exists.

11. Governance

- **Re-evaluation cadence.** The evidence tables are refreshed at least quarterly and whenever a model we use is materially updated or a new model becomes relevant.
- **Change control.** Routing table changes are made by the engineering team, recorded here with the evidence that motivated them, and reflected in this document's version number.
- **Measurement in production.** We monitor, per activity, the share of traffic routed, review pass rates, revision rates and latency, so a route that looks good on paper but performs worse in practice is caught and reverted. Two measurements are outstanding: the **output-review rewrite rate for pastoral guidance on the everyday tier**, which decides whether that row moves up a tier, and the **measured share of crisis-labelled traffic**, which turns Section 9's per-answer arithmetic into a real blended cost.

12. Limitations, stated plainly

- Benchmarks measure benchmarks. They correlate with real usefulness but are not the same thing; contamination and protocol differences exist. That is why we require more than one source and monitor production behaviour.
- The tier cut rests on one composite. The Artificial Analysis Intelligence Index is independent and runs every model in the same harness, which is exactly what we need for a grouping — but it is one number from one evaluator, and it is versioned: the same models scored 58.9 and 59.9 under v4.1 in July and 61 and 62 under v4.1.1 in August (S5, S35). We use it to group models, never to rank them within a group.
- Classification is probabilistic. A small fraction of messages will be routed to a model other than the ideal one. The Conscience makes this a quality question, never a safety question.
- Vendors change models under the same name, and can withdraw them (Section 10). Our evidence carries dates for exactly this reason, and the cadence in Section 11 exists to catch it.

Appendix A — Glossary

- **Artificial Conscience (AC).** Edenly's policy layer: classification, policy composition, output review and revision around every answer.
- **Activity.** A category of question used only for routing (Section 7.2).
- **Model tier.** A capability band on the independent composite index (Section 5). Not a plan.
- **Plan tier.** A paid subscription level, which meters usage in credits and never selects a model (Section 5.3).
- **Default model.** The writing model for every activity without a specific route.
- **Fallback chain.** The ordered list of models tried when a provider fails.
- **Vendor-reported / independent.** Whether a benchmark figure was published by the model's maker or measured by a third party.

Appendix B — Sources

All sources retrieved 1 September 2026. Publication dates as stated by the source.

ID	Source	Date
S1	DeepSeek, model card DeepSeek-V4-Flash (huggingface.co/deepseek-ai/DeepSeek-V4-Flash)	undated
S2	DeepSeek-V4 technical report, arXiv 2606.19348 v1	19 Jun 2026

ID	Source	Date
S3	DeepSeek API change log (api-docs.deepseek.com/updates)	24 Apr – 21 Aug 2026
S4	DeepSeek news: V4-Pro GA and new pricing (api-docs.deepseek.com/news/news260813)	13 Aug 2026
S5	OpenAI, "GPT-5.6: Frontier intelligence that scales with your ambition" (openai.com/index/gpt-5-6)	9 Jul 2026, upd. 21 Aug 2026
S6	OpenAI API model page, gpt-5.6-luna (developers.openai.com/api/docs/models/gpt-5.6-luna)	undated
S7	OpenAI API pricing (developers.openai.com/api/docs/pricing)	undated
S8	OpenAI, GPT-5.6 System Card (deploymentsafety.openai.com/gpt-5-6)	9 Jul 2026
S9	OpenAI, GPT-5.6 August Updates (cdn.openai.com/pdf/GPT_5_6_August_Updates.pdf)	6 Aug 2026
S11	Google DeepMind, Gemini 3.5 Flash model card (deepmind.google/models/model-cards/gemini-3-5-flash)	19 May 2026
S13	Google, Gemini API pricing (ai.google.dev/gemini-api/docs/pricing)	28 Aug 2026
S14	Google DeepMind, Gemini 3.7 Flash model card, including its Google-run comparison columns (deepmind.google/models/model-cards/gemini-3-7-flash)	13 Aug 2026
S15	Anthropic, "Introducing Claude Sonnet 5" (anthropic.com/news/claude-sonnet-5)	30 Jun 2026
S17	Anthropic platform documentation: pricing and models overview (platform.claude.com/docs)	undated
S19	alphaXiv rendering of the Claude Opus 5 System Card	submitted 24 Jul 2026
S22	xAI, Grok 4.6 Model Card, rev. 17 Aug 2026 (media.x.ai)	12 Aug 2026
S23	xAI developer documentation, Grok 4.6 (docs.x.ai/developers/grok-4-6)	undated
S24	Arena text and category leaderboards (arena.ai/leaderboard/text, .../creative-writing, .../longer-query, .../multi-turn)	27 Aug 2026
S25	Artificial Analysis model pages (artificialanalysis.ai/models/...)	undated
S26	Artificial Analysis evaluation pages: GPQA Diamond, Humanity's Last Exam, Global-MMLU-Lite	undated
S27	Epoch AI Benchmarking Hub, FrontierMath Tiers 1–3 (epoch.ai/benchmarks/frontiermath-tiers-1-3-v2)	v2 12 Jun 2026

ID	Source	Date
S29	Vals AI leaderboards (vals.ai/benchmarks/swebench, .../lcb)	30 Aug 2026
S30	EQ-Bench 4 official results (github.com/EQ-bench/eqbench4, results/elo_results.json)	undated
S35	Artificial Analysis model pages (artificialanalysis.ai/models/gpt-5-6-sol, .../claude-fable-5, .../gemini-3-7-flash)	retrieved 1 Sep 2026
S36	Vals AI model pages (vals.ai/models/openai_gpt-5.6-sol, .../anthropic_claude-fable-5)	evaluated 9 Jul 2026 (Sol) / 9 Jun 2026 (Fable 5)
S37	OpenAI API model page, gpt-5.6-sol (developers.openai.com/api/docs/models/gpt-5.6-sol)	undated
S38	Anthropic, "Claude Fable 5 and Claude Mythos 5" launch announcement (anthropic.com/news/claude-fable-5-mythos-5)	9 Jun 2026
S39	Anthropic, "Redeploying Claude Fable 5" (anthropic.com/news/redeploying-fable-5)	events 12 Jun – 1 Jul 2026
S40	Anthropic platform docs, Claude Fable 5 model reference (platform.claude.com/docs/en/models/fable-5)	undated
S41	Google, "Gemini 3.7 Flash: our most intelligent workhorse model" (blog.google/.../introducing-gemini-3-7-flash)	13 Aug 2026

Edenly Engineering Team