# 技术分析与战略建议：为高级报告生成场景选择 **Gemini 1.5 Pro** 与 **2.5 Pro**

## 第 **1** 节：执行摘要

本报告旨在为您的开发团队提供一份详尽的技术分析和战略建议，核心目标是在 Google 的两大旗舰多模态模型——Gemini 1.5 Pro 和 Gemini 2.5 Pro 之间做出明智抉择。此项决策将直接影响贵公司报告生成平台的核心能力、产出质量和运营成本，该平台旨在通过处理用户上传的多种格式文件来生成覆盖全球常用类型的各类报告。

分析的核心结论如下：

1. 文件格式支持并非决定性因素：经过对 Google AI Files API (https://ai.google.dev/api/files) 及其与两大模型兼容性的深入研究，报告明确指出，Gemini 1.5 Pro 和 Gemini 2.5 Pro 在支持处理的文件格式（MIME 类型）范围上几乎完全一致。两者均能原生处理包括图像、音频、视频及 PDF/纯文本文档在内的多种主流格式。因此，在两个模型之间进行选择的依据不应是其支持格式的广度。
2. 核心差异在于推理与编程能力：两个模型之间最根本的区别在于其底层架构和由此带来的性能差异。Gemini 2.5 Pro 引入了革命性的“思考模型”（Thinking Model）架构，赋予其在处理复杂问题前进行内部推理和规划的能力。这一能力在编程和代码分析任务中表现得尤为突出。基准测试数据显示，Gemini 2.5 Pro 在代码生成、代码编辑和代理式编码（Agentic Coding）等关键指标上，相较于前代模型实现了代际性的性能飞跃。
3. 决策的本质：性能与成本的战略权衡：Gemini 2.5 Pro 代表了当前最顶尖的性能，能够生成质量更高、技术更精确、逻辑更复杂的报告，尤其是在处理包含代码或需要深度逻辑推理的源文件时。然而，这种卓越性能也伴随着更高的 API 调用成本。与之相比，Gemini 1.5 Pro 虽然在复杂推理上稍逊一筹，但其本身依然是一个能力强大且极具成本效益的模型，足以胜任大量标准报告的生成任务。

最终建议：

鉴于您的应用场景明确要求生成覆盖“世界上常用的报告”，且核心任务与编程和多文件格式处理紧密相关，本报告明确建议采用 **Gemini 2.5 Pro** 作为核心模型。选择 Gemini 2.5 Pro 是对最终报告质量的一项战略性投资。其卓越的推理和编码能力将显著提升复杂报告

的准确性和深度，从而减少人工修正的需求，并能解锁更多高级报告类型的可能性。其带来的产品质量提升和长期效益，预计将超过其较高的单位成本。

## 第 2 节：Google AI Files API：开发者参考指南

为了确保您的开发团队能够高效、准确地实现文件上传功能，本节将详细阐述 Google AI Files API 的核心机制、操作限制以及最关键的文件格式支持列表。该 API 是您整个工作流程的入口，正确理解和使用它是项目成功的基础。

### API 的目的与工作流程

Google AI Files API 并非一个永久性的文件存储解决方案（如 Google Drive 或 Google Cloud Storage），而是一个专为 Gemini 模型设计的临时文件暂存区 [1]。其核心设计理念是将文件上传这一耗时操作与模型推理请求（即 prompt）解耦。您的系统架构——先上传文件获取 URI，再将 URI 加入队列等待后续处理——完美契合了这一设计 [2]。

这种机制的优势在于：

- 效率提升：大型文件可以预先上传，模型调用时仅需传递一个轻量级的 URI，避免了在每次请求中都嵌入庞大的 base64 编码数据。
- 资源复用：同一个上传的文件可以在其生命周期内被多次调用，用于不同的 prompt 请求，而无需重复上传 [3]。

### 关键操作限制

开发团队在设计队列处理和错误处理逻辑时，必须考虑以下硬性限制：

- 文件生命周期：上传至 Files API 的文件仅保留 **48** 小时。超过此期限，文件将被自动删除，相关的 URI 也会失效 [2]。这意味着您的处理队列必须确保在 48 小时内消费掉这些 URI，并应建立相应的重试和失败处理机制，以应对文件过期的情况。
- 存储配额：每个 Google Cloud 项目的总存储容量上限为 20 GB，单个文件的最大体积为 2 GB [2]。这些限制对于绝大多数文档和媒体文件来说是充足的，但对于超大视频或

数据集，可能需要进行预处理或分块。

支持上传的文件格式列表

为了实现客户端的文件上传筛选功能，以下表格详尽列出了 Google AI Files API 支持的所有文件格式，以 MIME 类型和常见文件扩展名进行分类。这是您开发团队可以直接使用的权威参考。

表 1：**Google AI Files API** 支持的文件格式

| 类别 | 支持的 MIME 类型 | 常见文件扩展名 |
| --- | --- | --- |
| 图像 **(Image)** | image/png | .png |
| | image/jpeg | .jpeg, .jpg |
| | image/webp | .webp |
| | image/heic | .heic |
| | image/heif | .heif |
| 音频 **(Audio)** | audio/wav | .wav |
| | audio/mp3 | .mp3 |
| | audio/aiff | .aiff |
| | audio/aac | .aac |
| | audio/ogg | .ogg |
| | audio/flac | .flac |
| 视频 **(Video)** | video/mp4 | .mp4 |

| | | |
|---|---|---|
| | video/mpeg | .mpeg |
| | video/mov | .mov |
| | video/avi | .avi |
| | video/x-flv | .flv |
| | video/mpg | .mpg |
| | video/webm | .webm |
| | video/wmv | .wmv |
| | video/3gpp | .3gpp |
| **纯文本/文档 (Plain Text/Documents)** | text/plain | .txt |
| | text/html | .html |
| | text/css | .css |
| | text/javascript | .js |
| | application/x-javascript | .js |
| | text/x-typescript | .ts |
| | application/x-typescript | .ts |
| | text/csv | .csv |
| | text/markdown | .md |
| | text/x-python | .py |
| | application/x-python-code | .py |

| | | |
|---|---|---|
| | application/json | .json |
| | text/xml | .xml |
| | application/rtf | .rtf |
| | text/rtf | .rtf |
| | application/pdf | .pdf |

数据来源: [5]

一个重要的实践细节是，存在一些文件类型（如 Microsoft Word 的 .docx）虽然可以通过 Files API 成功上传，但在提交给 Gemini 模型进行处理时，可能会返回“不支持的 MIME 类型”错误 [8]。这揭示了“可上传”与“可处理”之间的关键差异。因此，您的文件上传过滤器不仅应参考上表，还应以模型实际能够处理的格式为最终依据，这将在下一节中详细阐述。

## 第 3 节：多模态处理能力对比分析

本节直接回答您的核心问题：Gemini 1.5 Pro 和 2.5 Pro 在处理多格式文件方面有何差异，哪一个支持的格式更多，以及它们是否能处理所有 Files API 支持的格式。

### 原生多模态：一个共同的强大基础

首先需要明确的是，从 Gemini 1.0 开始，整个 Gemini 家族的模型都被设计为“原生多模态”（Natively Multimodal）[9]。这意味着模型从预训练阶段开始，就被设计用来同时理解和推理跨越文本、代码、图像、音频和视频等多种信息类型。这与早期模型将不同模态的处理模块（如独立的图像识别模型）“拼接”在一起的方式有着本质区别 [9]。

这种原生多模态的架构是 Gemini 模型能够广泛处理多种文件格式的根本原因，也解释了为何不同版本的模型在基础格式支持上表现出高度的一致性。

端到端文件格式支持矩阵

为了直观地比较 Files API 的上传能力与两个模型的实际处理能力，下表提供了一个端到端的支持矩阵。这对于理解整个工作流程中的潜在瓶颈至关重要。

表 2：端到端文件格式支持矩阵 **(Files API vs. Gemini** 模型**)**

| 类别 | MIME 类型 | Files API 上传支持 | Gemini 1.5 Pro 处理支持 | Gemini 2.5 Pro 处理支持 | 备注 |
|---|---|---|---|---|---|
| 图像 | image/png, image/jpeg, image/webp, image/heic, image/heif | 是 | 是 | 是 | 两模型均完全支持主流图像格式。 |
| 音频 | audio/wav, audio/mp3, audio/aiff, audio/aac, audio/ogg, audio/flac | 是 | 是 | 是 | 两模型均完全支持主流音频格式。 |
| 视频 | video/mp4, video/mov, video/avi, video/mpeg, etc. | 是 | 是 | 是 | 两模型均完全支持主流视频格式。 |
| 文档 | application/pdf | 是 | 是 | 是 | PDF 是两模型原生支持的核心文档格式。 |
| | text/plain | 是 | 是 | 是 | 纯文本是两模型原生支持的核心文档格式。 |
| | text/html, text/css, text/javascri | 是 | 是 | 是 | 这些格式被视为纯文本的变体，均可处 |

| | | | | | |
|---|---|---|---|---|---|
| | pt, text/csv, text/markdown, application/json, text/xml, text/rtf | | | | 理。 |
| | application/vnd.openxmlformats-officedocument.wordprocessingml.document (.docx) | 是 | 否 | 否 | 可上传，但模型无法直接处理。需预处理。 |
| | application/vnd.openxmlformats-officedocument.spreadsheetml.sheet (.xlsx) | 是 | 否 | 否 | 可上传，但模型无法直接处理。需预处理。 |
| | application/vnd.openxmlformats-officedocument.presentationml.presentation (.pptx) | 是 | 否 | 否 | 可上传，但模型无法直接处理。需预处理。 |

数据来源: [13]

分析与结论

从上表可以得出两个关键结论：

1. 文件格式支持能力基本对等：在所有原生支持的格式（图像、音频、视频、PDF、纯文本及其变体）上，Gemini 1.5 Pro 和 Gemini 2.5 Pro 的能力是完全相同的。没有任何证据表明其中一个模型比另一个支持更多的文件类型。因此，文件格式的覆盖范围不应

成为您选择模型的决定性因素。这一发现将分析的重点从兼容性转向了更深层次的性能和能力差异。

2. 文档处理的细微差别与架构要求：尽管您的目标是处理“世界上常用的报告”，但模型本身在“文档”这一模态上的原生支持是精确且有限的，主要集中在 application/pdf 和 text/plain [13]。对于 Microsoft Office (docx, xlsx, pptx) 等极其常见的商业报告格式，模型无法直接解析其内容。

这一发现对您的系统架构提出了一个明确的、不可或缺的要求：必须在调用 **Gemini API** 之前，实施一个服务器端的文档转换预处理步骤。您的系统需要能够检测上传文件的类型，如果是非原生支持的文档格式，则应自动将其转换为 PDF 或纯文本。例如，可以部署一个使用 LibreOffice 的无服务器函数（Serverless Function）或调用第三方文档转换服务来实现这一流程。若不建立此预处理管道，您的应用将无法处理大量常见的用户上传文件，从而严重影响其核心功能。

## 第 4 节：架构与基础模型差异

在确认了两个模型在文件格式支持上基本一致后，理解它们在底层架构上的根本差异，就成为做出正确选择的关键。这种差异解释了它们在性能、成本和能力上的不同表现，并直接关系到哪个模型更适合您复杂的报告生成任务。

### **Gemini 1.5 Pro**：基于专家混合（**MoE**）架构的规模化效率

Gemini 1.5 Pro 的架构核心是专家混合（**Mixture-of-Experts, MoE**）[14]。这是一种稀疏模型架构，其工作原理可以通俗地理解为：模型内部拥有大量专攻不同任务的“专家”子网络。当处理一个输入（例如一个词或一个图像块）时，系统会通过一个门控网络（gating network）智能地选择并仅激活一小部分最相关的“专家”来处理该输入。

这种设计的革命性优势在于：

- 计算效率：模型可以拥有巨大的总参数量（从而具备强大的知识和能力），但在处理任何单个任务时，实际参与计算的参数量却相对较小。这使得模型在保持高性能的同时，训练和推理的计算成本远低于同等规模的密集型模型 [14]。
- 卓越的长上下文处理：MoE 架构是 Gemini 1.5 Pro 能够高效处理高达 100 万甚至更长 token 上下文的关键技术之一 [14]。它使得模型在面对海量信息（如整部代码库或数百页

的文档）时，依然能保持高效的检索和理解能力。

可以认为，Gemini 1.5 Pro 的架构设计是一次在“信息处理规模”上的重大突破，其核心优势在于高效地处理和检索海量信息。

**Gemini 2.5 Pro**：基于**“**思考模型**”**架构的高级推理

Gemini 2.5 Pro 则在 1.5 Pro 的基础上引入了一个新的范式——**“**思考模型**”**（**Thinking Model**）[17]。这不仅仅是一次性能的迭代，而是一次架构层面的进化，标志着模型从“信息检索”向“问题解决”的转变。

“思考”架构的核心特征包括：

- 内部推理过程：在生成最终答案之前，模型会进行一个内部的、多步骤的推理过程。它会探索不同的解题策略，分解复杂问题，并评估中间步骤，这个过程类似于人类的“草稿”或“思维链”[17]。这个过程是模型原生能力的一部分，而非仅仅通过提示工程（prompt engineering）诱导产生。
- 可控的**“**思考预算**”**（**thinkingBudget**）：开发者可以通过 API 参数 thinkingBudget 来精确控制模型用于“思考”的计算资源（以 token 计）[21]。这是一个强大的控制杠杆：
  - 动态适应：当不设置预算时，模型会根据任务的复杂性自适应地决定思考量[17]。
  - 成本与质量的平衡：对于简单的任务，可以设置较低的预算（甚至在 Flash 模型中关闭思考）以降低延迟和成本。对于需要深度分析的复杂报告，可以增加预算，让模型投入更多计算资源以获得更准确、更深入的分析结果[21]。

Gemini 2.5 Pro 的架构是一次在“信息处理深度”上的飞跃。其核心优势在于对信息进行复杂、多步骤的分析、综合和创造性生成。

架构演进的战略意义

从 1.5 Pro 的 MoE 架构到 2.5 Pro 的“思考模型”架构，体现了 Google AI 模型的战略演进方向。1.5 Pro 解决了“如何高效读完一本大书”的问题，而 2.5 Pro 则致力于解决“读完这本书后，如何写出一篇深刻的书评”的问题。

对于您的报告生成平台而言，这一演进至关重要。您的应用不仅仅是简单地从上传的文件中提取或总结信息，而是要根据复杂的提示，对这些信息（尤其是代码）进行分析、比较、

评估，并最终生成一份结构化、逻辑严谨的全新报告。这种任务的性质与 Gemini 2.5 Pro 的“问题解决”导向架构高度契合，为其在下一节的基准测试中取得优异表现奠定了理论基础。

## 第 5 节：性能深入探讨：编程与推理基准测试

理论架构的先进性最终需要通过客观的性能数据来验证。本节将深入分析 Gemini 1.5 Pro 和 2.5 Pro 在编程和复杂推理任务上的量化表现，这些任务与您的报告生成应用场景直接相关。分析将基于行业公认的基准测试，为您提供选择模型的硬数据支持。

### 解读关键基准测试

为了准确评估模型的编程能力，我们选取了几个行业内最具挑战性和代表性的基准测试：

- **LiveCodeBench**：该测试评估模型解决类似编程竞赛问题的能力，考验其算法理解和代码生成的核心水平 [17]。
- **Aider Polyglot**：该测试衡量模型在现有代码库中编辑整个文件的能力，这对于分析和修改现有项目以生成报告至关重要 [17]。
- **SWE-bench**：这是衡量“代理式编码”（Agentic Coding）能力的黄金标准。它要求模型像一名软件工程师一样，自主地解决真实 GitHub 代码库中的实际问题（issues）。这项测试全面评估了模型理解问题、分析代码、制定计划并执行修复的能力，与您的端到端报告生成流程高度相似 [17]。

### 编程与推理能力正面对决

下表整合了来自官方技术报告和基准测试排行榜的数据，直观地展示了 Gemini 2.5 Pro 相对于前代及其他顶尖模型的性能优势。

表 3：编程与推理基准测试分数对比

| 基准测试 | 任务描述 | Gemini 1.5 Pro | Gemini 2.5 Pro | 性能差距 |
| --- | --- | --- | --- | --- |

| | | | | |
|---|---|---|---|---|
| | | (预估/参考) | (Thinking) | |
| **LiveCodeBench** | 代码生成 | 29.7% [1] | 69.0% | 巨大提升 |
| **Aider Polyglot** | 代码编辑 | N/A | 82.2% | 业界顶尖水平 |
| **SWE-bench Verified** (多轮尝试) | 代理式编码 | N/A | 67.2% | 业界顶尖水平 |
| **GPQA diamond** (科学) | 复杂科学推理 | N/A | 86.4% | 业界顶尖水平 |
| **AIME 2025** (数学) | 高级数学推理 | 17.5% [1] | 88.0% | 巨大提升 |

数据来源: [17]

[1] 注：由于最新的基准测试报告主要聚焦于 2.5 系列模型，直接与 1.5 Pro 在完全相同条件下的对比数据有限。此处 1.5 Pro 的分数引用自一份报告，该报告显示 2.5 Flash 模型（2.5 系列中性能较低的版本）在 LiveCodeBench 和 AIME 2025 上的得分（分别为 59.3% 和 72.0%）已数倍于 1.5 Pro 23。这有力地证明了从 1.5 Pro 到 2.5 Pro 的性能提升是跨越式的。

结果分析

数据清晰地揭示了几个不容忽视的事实：

1. 性能差距是代际性的，而非迭代性的：Gemini 2.5 Pro 在所有关键编程和推理基准上的表现都远超前代模型。在 LiveCodeBench 和 AIME 这类考验核心逻辑推理的任务上，性能提升不是百分之几，而是数倍的增长。这表明 2.5 Pro 的“思考模型”架构确实带来了根本性的能力突破。
2. 在实际编程任务中表现卓越：高达 82.2% 的 Aider Polyglot 分数意味着 Gemini 2.5 Pro 极擅长理解和修改复杂的现有代码，这对于从客户代码库生成分析报告或文档的任务来说是一项核心优势。
3. **“代理式编码”能力是真正的亮点：**在 SWE-bench 上取得的 67.2% 的高分是其最强有力的证明。这个基准测试模拟了从接收一个高层次需求（类似您的“报告提示”）到分析一个复杂的环境（上传的文件/代码库），再到执行一系列操作以达成目标的全过程。

Gemini 2.5 Pro 在此项测试中的领先地位，是其能够胜任您端到端报告生成工作流的最直接、最有力的证据。它表明该模型不仅能“写代码”，更能“解决问题”。

综上所述，纯粹从性能角度出发，尤其是在对编程能力要求极高的应用场景中，Gemini 2.5 Pro 相对于 1.5 Pro 拥有无可争议的、压倒性的优势。

## 第 6 节：运营与战略考量

除了核心性能，选择一个大型语言模型还需综合考量其在实际部署中的运营效率、成本效益和可扩展性。本节将分析这些关键的实际因素。

### 上下文窗口与长上下文性能

- 共享的优势：Gemini 1.5 Pro 和 2.5 Pro 都拥有一个非常巨大的 100 万 token 上下文窗口，并计划扩展到 200 万 token [16]。这是一个共同的、强大的优势，使得两个模型都非常适合需要一次性处理大量信息的任务，例如分析整个代码仓库、多份长篇文档或数小时的音视频记录。
- 窗口内的表现：拥有大窗口是一方面，在窗口内精准检索信息的能力同样重要。在“大海捞针”（Needle-in-a-Haystack）测试（如 MRCR v2 基准）中，Gemini 2.5 Pro 在 128k token 长度下表现出 58.0% 的召回率，优于 2.5 系列的其他模型，这表明其在长上下文中进行信息检索的可靠性更高 [17]。对于需要从大量文件中精确提取关键信息来生成报告的场景，这种高保真度的检索能力至关重要。

### 延迟与吞吐量

- 延迟的考量：Gemini 2.5 Pro 的“思考”过程是其强大推理能力的源泉，但这个过程需要消耗额外的计算时间和 token，因此不可避免地会引入比非思考模型更高的延迟 [28]。对于需要处理数十万 token 的复杂提示，响应时间可能达到数分钟 [30]。
- 架构的战略优势：您的异步、队列驱动的系统架构在这里显示出其战略价值。由于最终用户提交任务后无需同步等待结果，而是异步接收通知，因此模型的高延迟对用户体验的直接影响被降到最低。系统的整体吞吐量（单位时间内完成的报告数量）比单次请求的延迟更为重要。这种架构选择极大地降低了采用功能更强大但可能更慢的 2.5

Pro 模型的风险，使其高延迟成为一个可管理的运营成本，而非致命的用户体验缺陷。

经济分析：定价模型

成本是任何技术选型中的决定性因素。下表清晰地对比了两个模型的定价结构，以便进行直接的财务评估。

表 **4**：**API** 调用成本对比 **(**每 **100** 万 **tokens)**

| 模型 | 输入 Tokens 价格 (USD) | 输出 Tokens 价格 (USD) | 备注 |
| --- | --- | --- | --- |
| **Gemini 1.5 Pro** | $3.50 ( prompts < 128K) | $10.50 ( prompts < 128K) | 价格在 2024 年 10 月 1 日后有大幅下调 [31]。 |
| **Gemini 2.5 Pro** | $1.25 ( prompts ≤ 200K) $2.50 ( prompts > 200K) | $10.00 ( prompts ≤ 200K) $15.00 ( prompts > 200K) | 输出价格包含**"**思考**"**过程消耗的 **tokens**。 |

数据来源: [17]

从定价中可以观察到：

1. **"**思考**"**是有形成本：Gemini 2.5 Pro 的定价模型明确将"思考"过程的消耗计入输出 token 成本。这使得其高级推理能力成为一个可量化、可控制的开销。您的应用程序可以通过动态调整 thinkingBudget 参数，在任务复杂性与成本之间进行精细的平衡，这为成本优化提供了新的维度。
2. 输入成本优势：值得注意的是，对于中等长度（≤ 200K tokens）的提示，Gemini 2.5 Pro 的输入 token 价格实际上比 1.5 Pro 更低。考虑到您的应用需要处理大量文件内容作为输入，这可能会在一定程度上抵消其较高的输出成本。

最终的成本效益分析，需要将模型的性能优势（更高的报告质量、更低的失败率和修正成本）与 API 的直接费用进行综合权衡。对于高价值的复杂报告，2.5 Pro 带来的质量提升很可能使其具备更高的投资回报率。

## 第 **7** 节：最终建议与战略指南

综合以上对文件格式支持、底层架构、性能基准和运营成本的全面分析，本报告旨在为您提供一个明确、可执行的最终建议，以指导您的技术选型和后续开发。

## 核心结论回顾

- 文件格式支持已非瓶颈：Gemini 1.5 Pro 与 2.5 Pro 在多模态文件格式支持上能力对等，均能满足您处理多样化文件的需求。决策的关键不在于“能处理什么”，而在于“处理得怎么样”。
- 性能存在代际差异：Gemini 2.5 Pro 的“思考模型”架构为其带来了在复杂推理和编程任务上质的飞跃。基准测试数据无可辩驳地证明了其在代码生成、编辑和解决实际编程问题方面的绝对优势[17]。
- 成本与价值的权衡：2.5 Pro 的卓越性能伴随着更高的价格，但其定价结构和可控的 thinkingBudget 参数也为成本管理提供了灵活性[21]。

## 主要建议：优先采用 **Gemini 2.5 Pro**

本报告的最终建议是，将 **Gemini 2.5 Pro** 作为您报告生成平台的核心模型。

做出此项建议的理由如下：

1. 与核心价值主张高度契合：您的平台价值在于生成高质量、高复杂度的报告。选择在推理和编码能力上处于业界顶尖水平的 2.5 Pro，是确保产品核心竞争力的最直接方式。其性能优势并非微不足道的增量改进，而是能够直接提升最终产品质量的代际飞跃。
2. 投资于质量，降低隐性成本：虽然 2.5 Pro 的 API 调用成本更高，但一份由其生成的准确、深刻的报告所创造的价值，远超一份由次级模型生成的、需要大量人工审核和修正的报告。投资于更高质量的初始输出，可以显著降低下游的人工干预成本、提高用户满意度，并避免因报告错误导致的潜在商业风险。
3. 解锁未来可能性：2.5 Pro 强大的代理式编码和复杂推理能力，不仅能满足您当前的需求，更能为平台未来的功能扩展奠定坚实基础，例如生成更具交互性的代码分析报告、自动化的代码重构建议，或是更深度的多文档交叉分析报告。

次要建议：考虑混合模型策略作为长期优化

在项目初期，集中资源于 2.5 Pro 是最稳妥的策略。但随着业务的成熟和对成本优化的需求增加，可以考虑实施混合模型路由策略：

- 高端任务使用 **2.5 Pro**：对于需要深度代码分析、跨多份复杂文档进行推理等高价值、高难度的报告请求，继续使用 2.5 Pro，并根据需要调整 thinkingBudget 以确保最高质量。
- 标准任务使用 **2.5 Flash**：对于相对简单的任务，如总结单份纯文本文档、从结构化数据生成标准格式的报告等，可以将请求路由到成本更低、延迟更低的 Gemini 2.5 Flash 模型。该模型同样具备 100 万 token 上下文窗口和强大的基础能力，足以胜任这些任务，从而实现整体运营成本与性能的最佳平衡。

可执行的后续步骤

为将此战略建议转化为实际行动，建议您的团队按以下步骤推进：

1. 实施文件上传过滤器：根据本报告第 **2** 节的表 **1**，在应用前端和后端实施严格的文件类型和 MIME 类型校验，确保只有模型可处理的文件才能进入处理流程。
2. 构建文档预处理管道：立即启动一个关键子项目，开发一个服务器端的文档转换服务。该服务需能将 .docx、.xlsx、.pptx 等非原生支持的格式，可靠地转换为模型可以处理的 application/pdf 或 text/plain 格式。
3. 优先集成 **Gemini 2.5 Pro**：在核心开发工作中，直接基于 Gemini 2.5 Pro 的 API 进行集成。设计 API 调用逻辑时，要将 thinkingBudget 作为一个可配置参数，为未来的动态调整预留接口。
4. 进行小规模验证（**PoC**）：选取 3-5 个最具代表性的复杂报告生成场景，使用您自己的真实数据，分别调用 1.5 Pro、2.5 Flash 和 2.5 Pro 进行测试。通过对比产出报告的质量，来亲身验证本报告的结论，并为不同任务类型选择何种模型（或何种 thinkingBudget）积累第一手数据。
5. 设计灵活的路由架构：在系统设计层面，应将模型选择设计为可配置或可动态路由的。这将使您在未来能够无缝引入混合模型策略，或在 Google 推出新模型时能够快速切换，保持技术领先。

引用的著作

1. Release notes | Gemini API | Google AI for Developers, 访问时间为 七月 22, 2025, https://ai.google.dev/gemini-api/docs/changelog

2. Files API | Gemini API | Google AI for Developers, 访问时间为 七月 22, 2025, https://ai.google.dev/gemini-api/docs/files
3. Using files | Gemini API | Google AI for Developers, 访问时间为 七月 22, 2025, https://ai.google.dev/api/files
4. Using the Gemini File API for Prompts with Media - Raymond Camden, 访问时间为 七月 22, 2025, https://www.raymondcamden.com/2024/05/21/using-the-gemini-file-api-for-prompts-with-media
5. Supported input files and requirements | Firebase AI Logic - Google, 访问时间为 七月 22, 2025, https://firebase.google.com/docs/ai-logic/input-file-requirements
6. Gemini 2.5 Pro | Generative AI on Vertex AI - Google Cloud, 访问时间为 七月 22, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro
7. prompting_with_media.ipynb - Colab, 访问时间为 七月 22, 2025, https://colab.research.google.com/github/google/generative-ai-docs/blob/main/site/en/gemini-api/docs/prompting_with_media.ipynb
8. Does Gemini API Support all file mime types - Stack Overflow, 访问时间为 七月 22, 2025, https://stackoverflow.com/questions/78888864/does-gemini-api-support-all-file-mime-types
9. Introducing Gemini: our largest and most capable AI model - Google Blog, 访问时间为 七月 22, 2025, https://blog.google/technology/ai/google-gemini-ai/
10. Gemini 1 Report | PDF | Accuracy And Precision | Cognitive Science - Scribd, 访问时间为 七月 22, 2025, https://www.scribd.com/document/708909181/gemini-1-report
11. Release Notes: Gemini's multimodality - YouTube, 访问时间为 七月 22, 2025, https://www.youtube.com/watch?v=K4vXvaRV0dw
12. Google Launches Gemini, Its New Multimodal AI Model - Encord, 访问时间为 七月 22, 2025, https://encord.com/blog/gemini-google-ai-model/
13. Gemini 1.5 Pro | Generative AI on Vertex AI | Google Cloud, 访问时间为 七月 22, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-pro
14. Gemini 1.5 Pro - Prompt Engineering Guide, 访问时间为 七月 22, 2025, https://www.promptingguide.ai/models/gemini-pro
15. Gemini 1.5 Technical Report: Key Reveals and Insights - Gradient Flow, 访问时间为 七月 22, 2025, https://gradientflow.com/gemini-1-5-technical-report/
16. Gemini 1.5: Unlocking multimodal understanding across ... - arXiv, 访问时间为 七月 22, 2025, http://arxiv.org/pdf/2403.05530
17. Gemini - Google DeepMind, 访问时间为 七月 22, 2025, https://deepmind.google/models/gemini/
18. Gemini 2.5 Pro benchmarks released : r/singularity - Reddit, 访问时间为 七月 22, 2025, https://www.reddit.com/r/singularity/comments/1jjoeq6/gemini_25_pro_benchmarks_released/
19. Gemini 2.5 Pro - Google DeepMind, 访问时间为 七月 22, 2025,

https://deepmind.google/models/gemini/pro/
20. Google Gemini 2.5 Pro: The Thinking AI That's Redefining Intelligence - Medium, 访问时间为 七月 22, 2025, https://medium.com/@cognidownunder/google-gemini-2-5-pro-the-thinking-ai-thats-redefining-intelligence-22ce8c545e9a
21. Gemini thinking | Gemini API | Google AI for Developers, 访问时间为 七月 22, 2025, https://ai.google.dev/gemini-api/docs/thinking
22. Gemini 2.5: Our most intelligent models are getting even better, 访问时间为 七月 22, 2025, https://blog.google/technology/google-deepmind/google-gemini-updates-io-2025/
23. The most important takeaways from Google's Gemini 2.5 Paper - Devansh, 访问时间为 七月 22, 2025, https://machine-learning-made-simple.medium.com/the-most-important-takeaways-from-googles-gemini-2-5-paper-b43888c5cc65
24. Gemini 2.5 Pro: Features, Tests, Access, Benchmarks & More | DataCamp, 访问时间为 七月 22, 2025, https://www.datacamp.com/blog/gemini-2-5-pro
25. Gemini 2.5: Our most intelligent AI model - Google Blog, 访问时间为 七月 22, 2025, https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/
26. Google's Gemini 2.5 Pro model tops LMArena by close to 40 points - R&D World, 访问时间为 七月 22, 2025, https://www.rdworldonline.com/googles-gemini-2-5-pro-model-tops-lmarena-by-40-points-outperforms-competitors-in-scientific-reasoning/
27. Our next-generation model: Gemini 1.5 - Google Blog, 访问时间为 七月 22, 2025, https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/
28. Gemini 2.5 Pro, Thinking and Non-thinking - Google AI Developers Forum, 访问时间为 七月 22, 2025, https://discuss.ai.google.dev/t/gemini-2-5-pro-thinking-and-non-thinking/75269
29. Gemini 2.5 Pro - Intelligence, Performance & Price Analysis, 访问时间为 七月 22, 2025, https://artificialanalysis.ai/models/gemini-2-5-pro
30. Re: Gemini 2.5 Pro – Extremely High Latency on Large Prompts (100K–500K Tokens), 访问时间为 七月 22, 2025, https://www.googlecloudcommunity.com/gc/AI-ML/Gemini-2-5-Pro-Extremely-High-Latency-on-Large-Prompts-100K-500K/m-p/904130/highlight/true
31. Updated production-ready Gemini models, reduced 1.5 Pro pricing, increased rate limits, and more - Google Developers Blog, 访问时间为 七月 22, 2025, https://developers.googleblog.com/en/updated-gemini-models-reduced-15-pro-pricing-increased-rate-limits-and-more/
32. Gemini Developer API Pricing | Gemini API | Google AI for Developers, 访问时间为 七月 22, 2025, https://ai.google.dev/gemini-api/docs/pricing