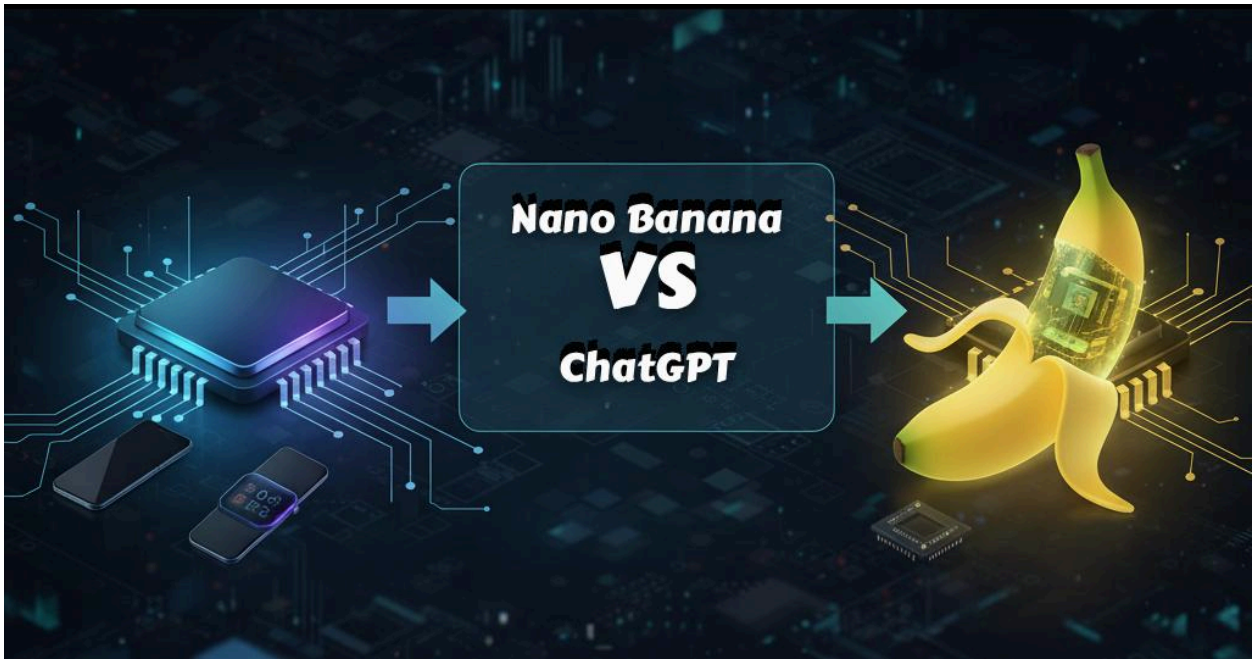
Gemini Nano και Nano Banana: Αρχιτεκτονική, Εφαρμογές και Ανταγωνιστική Ανάλυση Έναντι του Οικοσυστήματος ChatGPT



Κεφάλαιο 1: Διευκρίνιση Ορολογίας και Στρατηγική Θέση

Η εισαγωγή των μοντέλων Gemini Nano από την Google σηματοδοτεί μια στρατηγική στροφή προς την εκτέλεση της τεχνητής νοημοσύνης απευθείας στη συσκευή (*on-device*). Αυτό το κεφάλαιο εστιάζει στην αποσαφήνιση της ορολογίας, στην κατανόηση της τεχνικής βάσης (AICore) και στην ανάλυση της ραγδαίας εξέλιξης των μοντέλων Nano, θέτοντας τα θεμέλια για τη στρατηγική σύγκριση με τις cloud-based λύσεις όπως το ChatGPT.

1.1 Η Δυαδική Φύση του "Nano" της Google: Gemini Nano (SLM) vs. Nano Banana (Εικόνα)

Το Gemini Nano δεν είναι ένα ενιαίο προϊόν, αλλά μια στρατηγική ομπρέλα που καλύπτει δύο διακριτές κατηγορίες ικανοτήτων. Ο **Gemini Nano** είναι το Small Language Model (SLM) της Google, το οποίο είναι βελτιστοποιημένο για εκτέλεση απευθείας στον επεξεργαστή της συσκευής.[1] Ο πρωταρχικός του επιχειρησιακός ρόλος είναι η παροχή λειτουργιών Generative AI υψηλής συχνότητας με πολύ χαμηλή καθυστέρηση και υψηλή προστασία του απορρήτου (Privacy-First).[2]

Από την άλλη πλευρά, ο όρος **Nano Banana** φαίνεται να είναι μια εμπορική ονομασία [3], η οποία αναφέρεται στις προηγμένες ικανότητες επεξεργασίας και δημιουργίας εικόνας που ενσωματώνονται στα πολυτροπικά μοντέλα Gemini, και συγκεκριμένα στο

Gemini 2.5 Flash Image.[4] Ενώ η λειτουργία "Nano Banana" στοχεύει σε δημιουργικά αποτελέσματα και είναι πολυτροπική, ενσωματώνει λειτουργίες που πιθανώς απαιτούν τη συνδυαστική χρήση τοπικών (on-device) και cloud πόρων, ανάλογα με την πολυπλοκότητα της ζητούμενης εργασίας.[4] Η αρχιτεκτονική του Nano εξελίσσεται ραγδαία, με το Gemini Nano 2.0 να υποστηρίζει πλέον πολυτροπικότητα (Text + Images) [5], γεγονός που υποδηλώνει τη δυνατότητα να εκτελεί τοπικά εργασίες σχετικές με το Nano Banana, όπως η περιγραφή εικόνων ή η προ-επεξεργασία.[6]

1.2 Αρχιτεκτονική Πυρήνα: Λειτουργία του Gemini Nano μέσω του Android AICore

Το κλειδί για την επιτυχία του Gemini Nano ως λύσης on-device είναι η βαθιά του ενσωμάτωση στο λειτουργικό σύστημα Android, μέσω της υπηρεσίας **AICore**.[1] Το AICore λειτουργεί ως ένα κεντρικό σύστημα διαχείρισης μοντέλων (model manager) για on-device AI.

Αυτή η υπηρεσία επιτρέπει την αποτελεσματική αξιοποίηση του ειδικού hardware της συσκευής, όπως οι επεξεργαστές Google Tensor G4 που βρίσκονται στα Pixel 9, με αποτέλεσμα την επίτευξη χαμηλής καθυστέρησης συμπερασμάτων (*low inference latency*).[1] Επιπλέον, το AICore αναλαμβάνει κρίσιμες λειτουργίες, όπως η διατήρηση του μοντέλου ενημερωμένου (

keeps the model up-to-date) [8] και η παροχή ενσωματωμένων χαρακτηριστικών ασφαλείας, τα οποία αξιολογούν τις εξόδους του Gemini Nano έναντι των φίλτρων ασφαλείας της Google.[8]

Μία κρίσιμη στρατηγική συνέπεια της χρήσης του AICore είναι ότι επιτρέπει σε κάθε εφαρμογή να χρησιμοποιεί ένα **κοινόχρηστο μοντέλο Gemini Nano** που είναι ήδη εγκατεστημένο στη συσκευή.[9] Δεδομένου ότι τα Small Language Models (SLMs) μπορούν να έχουν μέγεθος γύρω στα 2 GB [5], η κοινή χρήση του μοντέλου αποφεύγει την ανάγκη για πολλαπλές λήψεις μεγάλων αρχείων, εξοικονομώντας σημαντικά τον αποθηκευτικό χώρο της συσκευής. Η Google παρέχει επίσης πρόσβαση στο

LiteRT-LM, τον κινητήρα συμπερασμάτων που τροφοδοτεί το Nano σε διάφορες πλατφόρμες, όπως Chrome, Chromebook Plus και Pixel Watch, επιτρέποντας την ανάπτυξη προσαρμοσμένων, υψηλής απόδοσης AI pipelines.[2]

1.3 Τεχνικές Προδιαγραφές και Εξέλιξη (Nano 1.0 σε 2.0)

Το Gemini Nano έχει υποστεί ραγδαία εξέλιξη, με βασικό στόχο τη βελτιστοποίηση των επιδόσεων σε κινητά περιβάλλοντα, όπου οι πόροι (μνήμη, ισχύς, μπαταρία) είναι περιορισμένοι. Η μετάβαση από το Nano 1.0 στο Nano 2.0 αποδεικνύει αυτή τη δέσμευση.

Η εξέλιξη αυτή φαίνεται ξεκάθαρα στη βελτίωση των βασικών προδιαγραφών:

Συγκριτικές Προδιαγραφές Μοντέλων Gemini Nano

Προδιαγραφή	**Gemini Nano 1.0**	**Gemini Nano 2.0**	**Βελτίωση (Ποσοστό)**
Κατανάλωση RAM	2.2 GB	1.8 GB	-18.2%
Ταχύτητα Συμπερασμάτων (tokens/sec)	32	58	+81.3%
Πολυτροπικότητα (Multimodality)	Μόνο Κείμενο	Κείμενο + Εικόνες	Ενισχυμένη
Επίπτωση	12%	7%	-41.7%

Μπαταρίας (ανά ώρα)			
Ακρίβεια (MMLU)	72.3%	78.9%	+6.6%

Η παρατηρούμενη βελτίωση της τάξης του 81.3% στην ταχύτητα συμπερασμάτων και η μείωση της επίπτωσης στην μπαταρία κατά 41.7% [5] δεν αποτελεί απλώς ένα μοντελο-κεντρικό επίτευγμα. Αντίθετα, η αύξηση της αποδοτικότητας είναι άρρηκτα συνδεδεμένη με την αυξημένη ικανότητα του ειδικού Google silicon (όπως ο Tensor G4) να εκτελεί τις λειτουργίες του Nano.[7] Αυτή η συζευγμένη βελτιστοποίηση hardware/software δημιουργεί μια ισχυρή εξάρτηση από το οικοσύστημα της Google. Αυτή η εξάρτηση λειτουργεί ως ένα

Hardware Lock-in, εξασφαλίζοντας ότι οι βέλτιστες on-device επιδόσεις και η καλύτερη ενεργειακή απόδοση παραμένουν αποκλειστικές στις συσκευές Pixel, παρέχοντας ένα ισχυρό ανταγωνιστικό πλεονέκτημα.

Επιπλέον, η μετάβαση από το μοντέλο 1.0 (μόνο κείμενο) στο 2.0 (Κείμενο + Εικόνες) [5] δείχνει ότι το Nano κλείνει τη λειτουργική ψαλίδα μεταξύ των καθαρά cloud-based λύσεων (όπως το αρχικό GPT) και των edge μοντέλων. Η ενσωμάτωση εικόνας στο on-device Nano 2.0 επιτρέπει στους προγραμματιστές να δημιουργούν πιο πλούσιες εφαρμογές με πλήρη προστασία δεδομένων, όπως η τοπική αναγνώριση περιεχομένου οθόνης ή η περιγραφή εικόνων.[6]

Κεφάλαιο 2: Μηχανισμοί Ενσωμάτωσης Εφαρμογών και Ανάπτυξη

Για τους προγραμματιστές, η πρόσβαση στις δυνατότητες του Gemini Nano καθορίζεται από δύο κύρια μοντέλα πρόσβασης. Η επιλογή του κατάλληλου μηχανισμού εξαρτάται από τη φάση ανάπτυξης και τον απαιτούμενο βαθμό ελέγχου.

2.1 Μοντέλα Πρόσβασης για Προγραμματιστές: AI Edge SDK και ML Kit GenAI APIs

Η Google προσφέρει δύο διακριτές οδούς για την ενσωμάτωση του Gemini Nano σε εφαρμογές Android:

1. **Google AI Edge SDK (Πειραματική Πρόσβαση):** Αυτό το SDK απευθύνεται σε προγραμματιστές που αναζητούν να δοκιμάσουν και να ενισχύσουν τις εφαρμογές τους με τις πιο πρόσφατες on-device AI δυνατότητες. Η χρήση του απαιτεί συμφωνία με τους όρους του προγράμματος πειραματικής πρόσβασης.[10]
2. **ML Kit GenAI APIs (Υψηλού Επιπέδου/Παραγωγής):** Το ML Kit παρέχει ένα interface υψηλού επιπέδου (*high-level interface*) το οποίο διευκολύνει την άμεση ενσωμάτωση δημοφιλών λειτουργιών, όπως η περίληψη, η διόρθωση, η επανεγγραφή και η περιγραφή εικόνων.[1]

Τα ML Kit GenAI APIs παρέχουν έτοιμη ποιότητα (*out-of-the-box quality*) για συνηθισμένες περιπτώσεις χρήσης.[9] Το σημαντικότερο πλεονέκτημα αυτού του μοντέλου είναι ότι λειτουργεί εξ ολοκλήρου στη συσκευή, εξαλείφοντας το πρόσθετο κόστος server για κάθε κλήση API.[9] Δεδομένου ότι τα APIs αυτά είναι χτισμένα πάνω στο AICore, οι εφαρμογές αξιοποιούν το κοινόχρηστο μοντέλο Nano που είναι ήδη στη συσκευή, αποφεύγοντας έτσι την ανάγκη λήψης μοντέλων και επιτυγχάνοντας σημαντική εξοικονόμηση αποθηκευτικού χώρου.[9]

2.2 Οδηγός Εφαρμογής: Χρήση του ML Kit GenAI API για βασικές λειτουργίες

Η διαδικασία ενσωμάτωσης του Gemini Nano μέσω του ML Kit για λειτουργίες όπως η περίληψη είναι εξαιρετικά απλοποιημένη, γεγονός που επιτρέπει ακόμη και σε προγραμματιστές χωρίς εξειδίκευση στο Machine Learning να ενσωματώσουν generative AI. Αυτή η στρατηγική επιτάχυνσης της υιοθέτησης, με την παροχή εύχρηστων APIs, διευρύνει το προβάδισμα της Google στην AI κινητή πλατφόρμα.

Τα βασικά βήματα υλοποίησης για Android API level 26+ περιλαμβάνουν [11]:

1. **Προσθήκη Εξάρτησης:** Εισαγωγή της κατάλληλης εξάρτησης, όπως για την περίληψη: implementation("com.google.mlkit:genai-summarization:1.0.0-beta1").[11]
2. **Δημιουργία Αντικειμένου:** Δημιουργία ενός αντικειμένου Summarizer.
3. **Εκτέλεση Inference:** Δημιουργία αιτήματος (Request) και εκτέλεση συμπερασμάτων (inference).[11]

Το ML Kit GenAI API παρέχει τόσο επιλογές **streaming** όσο και **non-streaming** για τη λήψη των αποτελεσμάτων.[9] Η επιλογή streaming API συνιστάται για περιπτώσεις όπου το αναμενόμενο αποτέλεσμα είναι μεγάλο (π.χ., περίληψη μεγάλου κειμένου), καθώς παρέχει

απαντήσεις σταδιακά, βελτιώνοντας έτσι την αίσθηση της απόκρισης (

perceived latency) για τον χρήστη.

2.3 Παραδείγματα Κώδικα και Ροής Εργασίας

Για την πρακτική εφαρμογή, η χρήση του ML Kit για on-device AI παρέχει σημαντικά οφέλη σε σχέση με τις cloud APIs, ιδίως όσον αφορά τη διαχείριση ασφάλειας. Ενώ για τις cloud APIs (π.χ., Gemini API) απαιτούνται βέλτιστες πρακτικές ασφαλείας, όπως η χρήση **server-side calls** για διατήρηση της εμπιστευτικότητας του API key, η χρήση **ephemeral tokens** για client-side πρόσβαση, ή η προσθήκη **περιορισμών** στο κλειδί [13], η on-device χρήση μέσω ML Kit/AICore δεν απαιτεί καμία διαχείριση API key. Αυτό απλοποιεί δραματικά την ανάπτυξη και εξαλείφει τον κίνδυνο διαρροής κλειδιού, καθώς όλη η επεξεργασία γίνεται τοπικά.[9]

Είναι κρίσιμο να σημειωθεί ότι το Gemini Nano απαιτεί συμβατές συσκευές που διαθέτουν την απαραίτητη hardware υποστήριξη AI (π.χ., Galaxy S25 Ultra, Pixel 9+). Οι εξομοιωτές δεν υποστηρίζουν ακόμη το Gemini Nano.[12] Αυτή η απαίτηση υλικού υπογραμμίζει την εξάρτηση της λύσης Nano από ειδικούς επιταχυντές AI, οι οποίοι είναι απαραίτητοι για την επίτευξη των απαιτούμενων ταχυτήτων και της ενεργειακής απόδοσης.

Κεφάλαιο 3: Περιπτώσεις Χρήσης του Gemini Nano (SLM)

Οι βασικές εφαρμογές του Gemini Nano επικεντρώνονται σε λειτουργίες υψηλής συχνότητας που επωφελούνται άμεσα από την on-device εκτέλεση, όπως η απομάκρυνση της εξάρτησης από το δίκτυο, η διασφάλιση του απορρήτου και η παροχή άμεσης απόκρισης.

3.1 Αυτοματοποιημένη Περίληψη (Summarization)

Η δυνατότητα περίληψης είναι μία από τις πρώτες και πιο σημαντικές εφαρμογές του Gemini Nano. Στις συσκευές Pixel, το Nano ενεργοποιεί τη λειτουργία "**Summarize in Recorder**".[7] Αυτή η λειτουργία επιτρέπει στον χρήστη να μετατρέπει ηχογραφημένες συνομιλίες, συνεντεύξεις ή διαλέξεις σε εύπεπτες περιλήψεις των κυριότερων σημείων. Το

κύριο όφελος είναι η δυνατότητα παράκαμψης της μακράς διαδικασίας μεταγραφής ή ακρόασης, καθώς και η διαθεσιμότητα της περίληψης

offline, χωρίς να απαιτείται σύνδεση στο διαδίκτυο ή κινητό δίκτυο.[7]

Πέρα από τις εφαρμογές για καταναλωτές, το Gemini υποστηρίζει και λειτουργίες γραφείου. Χρησιμοποιείται (συχνά σε υβριδική λειτουργία cloud/on-device) για την αυτόματη περίληψη απαντήσεων ανοιχτού κειμένου στο Google Forms, με τη δυνατότητα δημιουργίας περιλήψεων όταν υπάρχουν μεταξύ 3 και 200 απαντήσεις.[14]

3.2 Βελτίωση Κειμένου (Proofreading & Rewrite)

Το Nano υποστηρίζει εργασίες βελτίωσης κειμένου που απαιτούν γρήγορη και ιδιωτική επεξεργασία. Η λειτουργία "**Magic Compose**" στην εφαρμογή Google Messages (διαθέσιμη σε Pixel 6 και νεότερα) χρησιμοποιεί το Gemini Nano για να μεταμορφώνει το γραπτό κείμενο σε διαφορετικά στυλ, χωρίς την ανάγκη σύνδεσης στο internet.[7] Αυτή η ικανότητα εξαλείφει την "αστάθεια του δικτύου" ως παράγοντα στην εμπειρία του χρήστη (UX). Η AI μετατρέπεται σε ένα

πραγματικό βοηθητικό εργαλείο ενσωματωμένο στο λειτουργικό σύστημα (OS), και όχι σε μια εξαρτημένη εφαρμογή cloud.

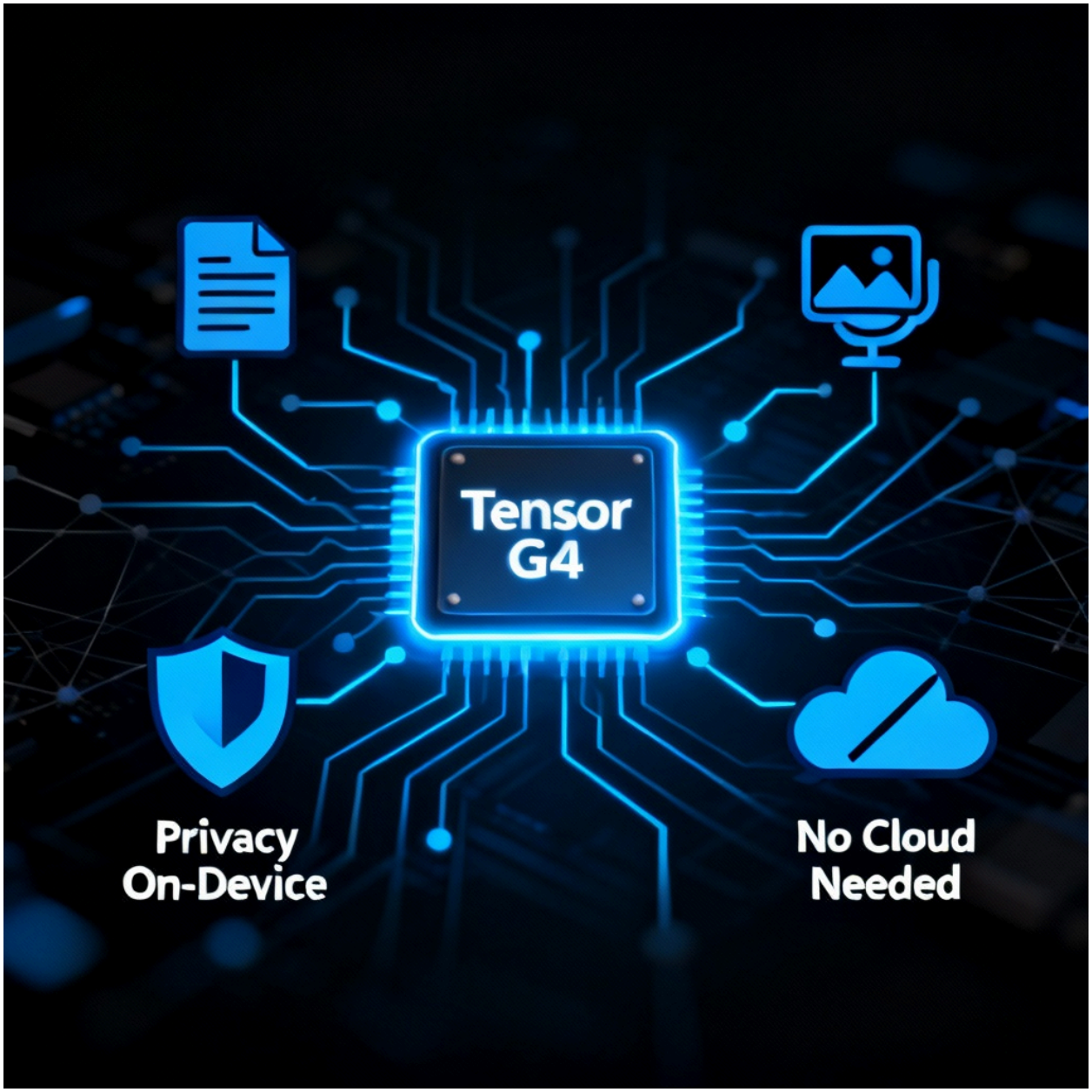
Επιπλέον, οι λειτουργίες **Proofreading** (διόρθωση γραμματικής και ορθογραφίας) και **Rewriting** (επανεγγραφή μηνυμάτων σε διαφορετικούς τόνους) είναι άμεσα προσβάσιμες μέσω του ML Kit GenAI API, επιτρέποντας στους προγραμματιστές να ενσωματώσουν αυτές τις βελτιώσεις σε σύντομα μηνύματα chat.[6]

3.3 On-Device Smart Reply

Μία κλασική και άκρως ευαίσθητη στην ιδιωτικότητα περίπτωση χρήσης είναι το Smart Reply. Το Nano χρησιμοποιείται για τη δημιουργία γρήγορων, συμφραζόμενων απαντήσεων (**Contextual Smart Reply**) σε εφαρμογές όπως το Gboard, το Gmail [15], το WhatsApp, το Line και το KakaoTalk.[16]

Το στρατηγικό πλεονέκτημα αυτής της λειτουργίας έγκειται στην προστασία των προσωπικών συνομιλιών. Δεδομένου ότι τα δεδομένα εισόδου, τα συμπεράσματα και τα δεδομένα εξόδου επεξεργάζονται **αποκλειστικά τοπικά** [9], το περιεχόμενο των μηνυμάτων

ή των emails δεν αποστέλλεται ποτέ στο cloud.[1] Αυτό είναι κρίσιμο για τις περιπτώσεις χρήσης όπου η χαμηλή καθυστέρηση και οι διασφαλίσεις απορρήτου είναι πρωταρχικής σημασίας.[1]



Κεφάλαιο 4: Εφαρμογή "Nano Banana" (Image Editing and Generation)

Η λειτουργία "Nano Banana" αντιπροσωπεύει τη φιλοδοξία της Google να πρωτοπορήσει στην πολυτροπικότητα, όχι μόνο στο cloud, αλλά και με επέκταση στις on-device δυνατότητες.

4.1 Το Nano Banana ως Μέρος του Gemini 2.5 Flash Image

Το Gemini αναπτύχθηκε αρχιτεκτονικά για να χειρίζεται πολυτροπικές εισόδους (κείμενο, εικόνες, κώδικας) από τη γέννησή του, σε αντίθεση με το ChatGPT/GPT, το οποίο πρόσθεσε αυτές τις δυνατότητες εκ των υστέρων.[4] Η λειτουργία "Nano Banana", ως εμπορικός όρος, εντάσσεται στην ευρύτερη ικανότητα επεξεργασίας εικόνας του

Gemini 2.5 Flash Image [4], το οποίο ανταγωνίζεται άμεσα την ικανότητα δημιουργίας εικόνας του ChatGPT μέσω του DALL·E.[4]

4.2 Εφαρμοσμένα Παραδείγματα (Creative Examples)

Οι δυνατότητες του Nano Banana (Gemini 2.5 Flash Image) επεκτείνονται πέρα από την απλή δημιουργία εικόνων, εστιάζοντας στη σύνθετη επεξεργασία και τη διατήρηση της συνοχής του θέματος:

1. **Μετατροπή Στυλ και Αντικειμένων:** Η ικανότητα να μετατρέψει ένα υπάρχον αντικείμενο, όπως ένα ψαλίδι, σε έναν φανταστικό χαρακτήρα, ή να αλλάξει το υλικό ενός ρούχου (π.χ., φόρεμα φτιαγμένο από μπάλες του τένις).[3]
2. **Τροποποίηση Σκηνικού και Αρχιτεκτονικής:** Η δυνατότητα να μετασχηματίσει ριζικά ένα περιβάλλον, όπως η μετατροπή μιας λευκής οικίας με κολώνες σε έναν ζωντανό τροπικό νησιωτικό σχεδιασμό, αντικαθιστώντας τη στέγη με άχυρο και προσθέτοντας μπαμπού.[3]
3. **Σύνθετες Αφηγήσεις Εικόνων:** Η δημιουργία εκτενών, συναρπαστικών ιστοριών που αποτελούνται από μια σειρά εικόνων (π.χ., 9 μέρη), οι οποίες διηγούνται μια ιστορία αποκλειστικά μέσω της οπτικής αναπαράστασης, χωρίς την προσθήκη κειμένου.[3]
4. **3D Μοντελοποίηση:** Η παραγωγή ρεαλιστικών 3D μοντέλων ενός ζώου από μια φωτογραφία, με δυνατότητα τοποθέτησής τους σε ένα συγκεκριμένο, ρεαλιστικό περιβάλλον (π.χ., δίπλα σε συσκευασία δώρου).[3]

4.3 Πολυτροπικότητα στην Πράξη: Διαδικασία Εισόδου/Εξόδου (Input/Output)

Ένα σημαντικό διαφοροποιητικό στοιχείο του Nano Banana (Gemini 2.5 Flash Image) είναι η ικανότητά του στη **συνεπή επεξεργασία χαρακτήρων και τη διατήρηση της σκηνής**.[4] Αυτό είναι ζωτικής σημασίας για επαγγελματικές ροές εργασίας, όπου οι χρήστες χρειάζονται επαναληπτικές επεξεργασίες, διατηρώντας την οπτική ταυτότητα του θέματος. Ενώ το DALL·E υπερέχει στη δημιουργική φαντασία, το Nano Banana φαίνεται να στοχεύει σε

επαναληπτικές διαδικασίες σχεδιασμού και γρήγορα mockups, όπου η συνέπεια έναντι του αρχικού θέματος είναι απαραίτητη.

Παρόλο που το Nano 2.0 υποστηρίζει πολυτροπικότητα [5], οι σύνθετες εργασίες που περιγράφονται (π.χ., 3D μοντελοποίηση) είναι εξαιρετικά απαιτητικές σε υπολογιστική ισχύ.[3] Ως εκ τούτου, η στρατηγική εφαρμογής είναι πιθανώς

υβριδική: το Nano on-device χειρίζεται τις αρχικές εισόδους, την περιγραφή και τις απλές διορθώσεις, ενώ το Flash Image (Nano Banana) στο cloud αναλαμβάνει τη βαριά γενετική επεξεργασία. Αυτή η κατανομή φόρτου διασφαλίζει ότι η Google μπορεί να προσφέρει τα οφέλη της χαμηλής καθυστέρησης για τις αρχικές ενέργειες, διατηρώντας παράλληλα την ποιότητα και την πολυπλοκότητα των τελικών αποτελεσμάτων.

Κεφάλαιο 5: Παραδείγματα Prompts και Πρακτική Εφαρμογή

Αυτό το κεφάλαιο παρέχει πρακτικά παραδείγματα (prompts) για τις δημιουργικές ικανότητες του **Nano Banana** (Gemini 2.5 Flash Image) και για τις λειτουργίες **Gemini Nano** (On-Device SLM).

5.1 Παραδείγματα Prompts για το Nano Banana (Δημιουργία & Επεξεργασία Εικόνας)

Το **Nano Banana** περιγράφει τις προηγμένες πολυτροπικές δυνατότητες επεξεργασίας εικόνας του Gemini 2.5 Flash, οι οποίες μπορούν να δοκιμαστούν στην εφαρμογή Gemini, σύμφωνα με τη Google.[3]

Κατηγορία	Prompt (Οδηγία)	Στόχος
Μετατροπή Αντικειμένου	«Μετέτρεψε αυτά τα ψαλίδια σε έναν ρεαλιστικό φανταστικό χαρακτήρα σε μια ταινία για ξωτικά και νεράιδες.»	Δημιουργία φανταστικού χαρακτήρα από καθημερινό αντικείμενο.[3]
Μετασχηματισμός Σκηνικού	«Μεταμόρφωσε αυτό το σπίτι σε έναν ζωντανό σχεδιασμό τροπικού νησιού. Αντικατάστησε τη στέγη με άχυρο και πρόσθεσε δομικά στοιχεία από μπαμπού. Περίβαλέ το με πλούσια, πολύχρωμα τροπικά φυτά και φοίνικες.»	Ριζική αλλαγή αρχιτεκτονικού στυλ και περιβάλλοντος.[3]
Δημιουργία 3D Μοντέλου	«Αριστερά: Ένα σκυλί Cavalier King Charles Spaniel ξαπλωμένο. Δεξιά: Prompt: Δημιούργησε ένα ρεαλιστικό μικρό 3D μοντέλο αυτού του σκύλου. Τοποθέτησε το μοντέλο σε ένα γραφείο δίπλα σε συσκευασία γενεθλίων που να δείχνει ότι κάποιος μόλις το ξετύλιξε ως δώρο.»	Σύνθετη τρισδιάστατη μοντελοποίηση και ενσωμάτωση σε συγκεκριμένη ρεαλιστική σκηνή.[3]
Διατήρηση/Αλλαγή Υλικού	«Άλλαξε το φόρεμα αυτού του ατόμου ώστε να είναι	Τροποποίηση υλικού ενός ρούχου διατηρώντας το

	φτιαγμένο από μπάλες του τένις.»	σχήμα.[3]
Αφήγηση Εικόνων (Visual Storytelling)	«Δημιούργησε μια συναρπαστική επική ιστορία 9 μερών με 9 εικόνες με αυτούς τους δύο πρωταγωνιστές και τις περιπέτειές τους ως μυστικοί υπερήρωες. Η ιστορία είναι thrilling throughout with emotional highs and lows and ending on a great twist and high note. Μην συμπεριλάβεις λέξεις ή κείμενο στις εικόνες, αλλά διηγήσου την ιστορία αποκλειστικά μέσω της εικονογράφησης.»	Δημιουργία μιας σειράς συνεπών εικόνων για σύνθετη αφήγηση.[3]

Χρήσιμος Σύνδεσμος	Περιγραφή
https://blog.google/products/gemini/gemini-nano-banana-examples/ [3]	Δείχνει παραδείγματα του Nano Banana και αναφέρει ότι η δυνατότητα μπορεί να δοκιμαστεί στην εφαρμογή Gemini.

5.2 Παραδείγματα Χρήσης για το Gemini Nano (On-Device SLM)

Το Gemini Nano (SLM) χρησιμοποιείται για **on-device** λειτουργίες [1], οι οποίες είναι ενσωματωμένες σε εφαρμογές μέσω των

ML Kit GenAI APIs.[9] Αυτές οι λειτουργίες δεν απαιτούν τη σύνταξη ενός prompt από τον χρήστη, αλλά ενεργοποιούνται αυτόματα βάσει του περιεχομένου που επεξεργάζεται η

συσκευή.

Λειτουργία	Παράδειγμα Prompt / Είσοδος (Input)	Αποτέλεσμα (Output)
Περίληψη (Summarization) (Μέσω ML Kit API)	**Είσοδος:** «Γεια σου Γιάννη, απλώς μια γρήγορη υπενθύμιση ότι έχουμε την παρουσίαση του προϊόντος στον πελάτη αύριο στις 10:00 π.μ. Βεβαιώσου ότι έχεις ενημερώσει το pitch deck και έχεις ελέγξει ξανά τα δεδομένα από την περασμένη εβδομάδα. Επίσης, η ομάδα UX είχε κάποιες σημειώσεις για το πρωτότυπο που ίσως αξίζει να αναφερθούν.»	**Έξοδος (Περίληψη):** «Υπενθύμιση: Παρουσίαση πελάτη αύριο στις 10 π.μ. Ενημέρωσε το pitch deck και έλεγξε τις σημειώσεις UX.» [12]
Διόρθωση/Επανεγγραφή (Proofreading / Rewrite) (Gboard/Messages)	**Είσοδος:** "Τον σινημα ειμασται. Να ρθεις?"	**Έξοδος (Διόρθωση):** "Στο σινεμά είμαστε. Να έρθεις;"
Smart Reply (Gboard, WhatsApp)	**Είσοδος (Μήνυμα):** "Έφτασα στο αεροδρόμιο, αλλά η πτήση μου καθυστέρησε 2 ώρες λόγω καιρού. Θέλεις να κανονίσουμε να με πάρεις αργότερα;"	**Έξοδος (Προτεινόμενες Απαντήσεις):** 1. "Ναι, κανένα πρόβλημα!" 2. "Εντάξει, στείλε μου μήνυμα όταν προσγειωθείς." 3. "Πόση ώρα καθυστέρησης; Τι ώρα θα έρθεις;" [15]

Χρήσιμος Σύνδεσμος	Περιγραφή
https://developer.android.com/ai/gemini-nano [1]	Επίσημη πηγή για το Gemini Nano και την πρόσβαση μέσω ML Kit GenAI APIs και AICore.

https://developer.android.com/ai/gemini-nano/ml-kit-genai [6]	Οδηγός για τους προγραμματιστές σχετικά με τις λειτουργίες Περίληψης, Διόρθωσης και Επανεγγραφής μέσω ML Kit.

Κεφάλαιο 6: Στρατηγική Σύγκριση: Gemini Nano (On-Device) vs. Cloud LLMs (ChatGPT)

Η στρατηγική διαφοροποίηση μεταξύ του Gemini Nano και των παραδοσιακών cloud-based μοντέλων, όπως το ChatGPT (GPT-4/5), βασίζεται σε τρεις άξονες: απόρρητο/ασφάλεια, οικονομικό κόστος λειτουργίας (TCO) και συμβιβασμοί επιδόσεων.

6.1 Το Στρατηγικό Πλεονέκτημα του On-Device: Απόρρητο, Ασφάλεια και Ανεξαρτησία

Το βασικό πλεονέκτημα του Gemini Nano είναι η αρχιτεκτονική του που εστιάζει στην προστασία των δεδομένων. Δεδομένου ότι τα δεδομένα εισόδου, τα συμπεράσματα και τα δεδομένα εξόδου επεξεργάζονται **τοπικά** [1], ευαίσθητες πληροφορίες (όπως προσωπικά μηνύματα ή ηχογραφήσεις) δεν αποστέλλονται ποτέ εκτός της συσκευής, εξασφαλίζοντας πλήρη ιδιωτικότητα.[18] Αντίθετα, η OpenAI διατηρεί συνήθως τις εισόδους API για 30 ημέρες, εκτός εάν ο χρήστης επιλέξει την εξαίρεση σε enterprise πλάνα.[4]

Όσον αφορά την ασφάλεια, η τοπική εκτέλεση μειώνει τον κίνδυνο διαρροών δεδομένων που συνδέονται με εξωτερικά δίκτυα.[19] Στις συσκευές Pixel, η προστασία ενισχύεται από το

Google Tensor G4 chip και το πιστοποιημένο **Titan M2** security chip, το οποίο παρέχει πολλαπλά επίπεδα hardware ασφαλείας.

Τέλος, η **λειτουργική ανεξαρτησία** του Nano είναι στρατηγικής σημασίας. Η λειτουργικότητα παραμένει 100% αμετάβλητη ακόμη και χωρίς αξιόπιστη σύνδεση internet.[7] Αυτό μετατρέπει την AI από μια ασταθή υπηρεσία cloud σε ένα αξιόπιστο εργαλείο ενσωματωμένο στο λειτουργικό σύστημα, γεγονός που την καθιστά ανώτερη σε συνθήκες χαμηλής ή μηδενικής κάλυψης δικτύου.

6.2 Οικονομική Ανάλυση: Μηδενικό Κόστος API του Nano έναντι Τιμολόγησης Token του ChatGPT (Cloud)

Η οικονομική ανάλυση αποκαλύπτει μια θεμελιώδη διαφορά στο μοντέλο κόστους.

Το Gemini Nano έχει ουσιαστικά **μηδενικό κόστος ανά κλήση API** (*zero per-API-call costs*).[2] Το κόστος του μοντέλου μετατίθεται πλήρως στο αρχικό, σταθερό κόστος αγοράς του hardware. Για εφαρμογές υψηλής συχνότητας (π.χ., εκατομμύρια smart replies ή διορθώσεις κειμένου ημερησίως), αυτό οδηγεί σε δραματική μείωση του Συνολικού Κόστους Ιδιοκτησίας (TCO).

Αντίθετα, τα μοντέλα της OpenAI (GPT) χρεώνουν βάσει του αριθμού των tokens (tokens εισόδου και εξόδου).[20] Για παράδειγμα, το GPT-5 (cloud) κοστολογείται $1.25 (Είσοδος) και $10.00 (Έξοδος) ανά 1 εκατομμύριο tokens. Ακόμη και το πιο οικονομικό μοντέλο cloud, το GPT-5 Nano, κοστολογείται $0.05 (Είσοδος) και $0.40 (Έξοδος) ανά 1 εκατομμύριο tokens.[20] Όταν ο όγκος των κλήσεων είναι τεράστιος, το συσσωρευμένο μεταβλητό κόστος των token APIs μπορεί γρήγορα να καταστήσει τη cloud λύση ασύμφορη, ενισχύοντας την οικονομική υπεροχή του Nano για επαναλαμβανόμενες ρουτίνες.

6.3 Επιδόσεις: Latency, Ακρίβεια και Ενεργειακή Κατανάλωση

Υπάρχουν σαφείς συμβιβασμοί μεταξύ των δύο αρχιτεκτονικών στον τομέα των επιδόσεων:

- **Latency (Καθυστέρηση):** Ενώ το Nano είναι εξαιρετικά βελτιστοποιημένο (π.χ., μέσω του κινητήρα LiteRT-LM), για σύνθετες εργασίες, η τοπική εκτέλεση σε περιορισμένο hardware μπορεί να έχει σημαντική καθυστέρηση (π.χ., >30 δευτερόλεπτα για ουσιαστική έξοδο [1]). Αντίθετα, οι cloud υπηρεσίες επιδεικνύουν καλύτερη χρονική απόδοση σε σύνθετες εργασίες (<10 δευτερόλεπτα [1]). Ωστόσο, για μικρό περιεχόμενο (100 λέξεις), η καθυστέρηση του on-device (15.07 s) είναι συγκρίσιμη με το remote (8.90 s).[8] Το στρατηγικό πλεονέκτημα του Nano είναι ότι εξαλείφει τη **δικτυακή καθυστέρηση (network latency)**, κάνοντας το μοντέλο να φαίνεται πιο αξιόπιστο και γρήγορο στην πράξη, καθώς ο χρόνος-μέχρι-το-πρώτο-token (TTFT) είναι εξαιρετικά χαμηλός.[2]
- **Ακρίβεια:** Τα μεγάλα cloud μοντέλα, όπως το GPT-4.1, διατηρούν υψηλότερη ακρίβεια (π.χ., 95.4% σε ορισμένα benchmarks) από τα βελτιστοποιημένα on-device SLMs (π.χ., Nano ~79.2% σε συγκρίσιμα σενάρια).[21] Αυτό σημαίνει ότι για κρίσιμες εργασίες που

απαιτούν την υψηλότερη δυνατή συλλογιστική ικανότητα (π.χ., νομική ανάλυση), τα cloud μοντέλα παραμένουν απαραίτητα.

- **Ενέργεια:** Η ενεργειακή βελτίωση του Nano, με μείωση της επίπτωσης μπαταρίας κατά 41.7% [5], καθιστά τη λύση on-device στρατηγικά ανώτερη για την παράταση της διάρκειας ζωής της μπαταρίας σε κινητά. Παρόλο που η εκτέλεση στο cloud είναι ενεργειακά απαιτητική στα κέντρα δεδομένων [22], η βελτιστοποίηση της ενεργειακής κατανάλωσης του Nano το καθιστά βιώσιμο για χρήση μεγάλης διάρκειας στη συσκευή.

Στρατηγική Σύγκριση: On-Device (Gemini Nano) vs. Cloud (GPT/ChatGPT)

Κριτήριο	**Google Gemini Nano (On-Device)**	**OpenAI GPT-4/5 (Cloud API)**
Τρόπος Λειτουργίας	Αποκλειστικά στη συσκευή (AICore/Tensor Chip)	Απομακρυσμένη επεξεργασία (Cloud Servers)
Κόστος API	Μηδέν (Σταθερό κόστος Hardware)	Βάσει tokens (Μεταβλητό κόστος)
Απόρρητο Δεδομένων	Υψηλό. Δεδομένα επεξεργάζονται τοπικά	Μεσαίο/Χαμηλό. Δεδομένα διατηρούνται 30 ημέρες (εκτός Enterprise opt-out)
Διαθεσιμότητα	100% Offline	Απαιτείται σταθερή σύνδεση δικτύου
Στόχος Εφαρμογής	Υψηλής συχνότητας, ιδιωτικές, μικρές εργασίες	Σύνθετη συλλογιστική (Reasoning), μεγάλες εργασίες, κώδικας

Κεφάλαιο 7: Βαθιά Ανάλυση: Gemini 2.5 Flash-Lite έναντι GPT-5 Nano (Budget Cloud Models)

Εκτός από τη μάχη on-device έναντι cloud, υπάρχει έντονος ανταγωνισμός στην κατηγορία των οικονομικά αποδοτικών cloud μοντέλων, τα οποία αναφέρονται ως "Budget Thinking"

Models. Αυτά τα μοντέλα αποτελούν μια γέφυρα μεταξύ της ταχύτητας και του κόστους.

7.1 Η Κατηγορία "Budget Thinking" Models

Η κατηγορία αυτή περιλαμβάνει μοντέλα που σχεδιάστηκαν για να προσφέρουν ισορροπία μεταξύ τιμής, ταχύτητας και αξιοπρεπούς ικανότητας συλλογιστικής.[21] Βασικοί ανταγωνιστές είναι το

Gemini 2.5 Flash-Lite (η πιο cost-efficient έκδοση του Gemini 2.5) και το **GPT-5 Nano** (με δυνατότητα ρύθμισης του reasoning_effort).[21]

7.2 Σύγκριση Επιδόσεων Υπό Φόρτο (Accuracy vs. Latency Trade-off)

Οι δύο εταιρείες έχουν υιοθετήσει διαφορετικές στρατηγικές για αυτήν την κατηγορία, προσφέροντας διαφορετικούς συμβιβασμούς μεταξύ ταχύτητας και γνωστικής ανθεκτικότητας (cognitive resilience).

Το **Flash-Lite** είναι ο ηγέτης στην ταχύτητα (*speed leader*) όταν ο γνωστικός φόρτος είναι χαμηλός και είναι βελτιστοποιημένο για υψηλό throughput.[21] Ωστόσο, η ακρίβειά του μειώνεται απότομα καθώς αυξάνεται η πυκνότητα των οδηγιών. Αντίθετα, το

GPT-5 Nano διατηρεί καλύτερη ακρίβεια (*better adherence*) υπό βαρύτερο γνωστικό φόρτο, λειτουργώντας ως ένα πιο ανθεκτικό μοντέλο "γέφυρα", αν και είναι πιο αργό.[21]

Συγκεκριμένα, σε δοκιμές με 100 οδηγίες:

- Το Flash-Lite επιτυγχάνει ταχύτητα ~20.4 s με ακρίβεια ~81.4%.
- Το GPT-5 Nano επιτυγχάνει ακρίβεια ~79.2% (συγκρίσιμη) αλλά είναι σημαντικά πιο αργό, με καθυστέρηση ~53.1 s.

Ωστόσο, σε σενάρια με εξαιρετικά υψηλή πυκνότητα οδηγιών (250 instructions), το GPT-5 Nano διατηρεί την ακρίβεια στο 44.7% έναντι του 33.0% του Flash-Lite, αποδεικνύοντας μεγαλύτερο γνωστικό περιθώριο, αν και η χρησιμότητα και των δύο μοντέλων μειώνεται δραστικά σε αυτό το επίπεδο φόρτου.[21]

Ανάλυση Επιδόσεων Μοντέλων Budget Cloud (100 Οδηγίες)

Μοντέλο	Ακρίβεια (Approx.)	Καθυστέρηση (Latency - Approx.)	Τιμή Εισόδου (ανά 1M tokens)	Στρατηγική Θέση
Gemini 2.5 Flash-Lite	81.4%	20.4 s	$0.10 [21]	Ταχύτητα και Μέγιστο Throughput
GPT-5 Nano	79.2%	53.1 s	$0.05 [21]	Γνωστική Ανθεκτικότητα / Bridge Model

Σημείωση: Οι τιμές εξόδου (Output) είναι ίδιες στα $0.40/1M tokens.[21]

7.3 Οικονομική Διαφοροποίηση και Context Window

Παρά την ομοιότητά τους, υπάρχουν βασικές οικονομικές και λειτουργικές διαφορές. Το GPT-5 Nano είναι φθηνότερο στην είσοδο, με κόστος $0.05 ανά 1M tokens σε σύγκριση με το $0.10 του Flash-Lite.[21] Αυτό σημαίνει ότι για εφαρμογές με μεγάλη είσοδο κειμένου (π.χ., ανάλυση μεγάλων εγγράφων) χωρίς απαίτηση για σύνθετη έξοδο, το GPT-5 Nano μπορεί να είναι η πιο οικονομική επιλογή.

Όσον αφορά το context window (το μέγιστο μήκος της εισόδου), το Gemini 2.5 Flash-Lite υποστηρίζει ένα τεράστιο παράθυρο 1M tokens [4], ενώ το GPT-5 Nano υποστηρίζει 400,000 tokens.[4] Το μεγάλο context window του Flash-Lite είναι καθοριστικό για την επεξεργασία πολύ μεγάλων εγγράφων.

7.4 Στρατηγική Κατάταξη: Πότε υπερτερεί το Flash-Lite και πότε το GPT-5 Nano

Η ανάλυση των budget cloud μοντέλων επιτρέπει τη χάραξη συγκεκριμένων **Microservices Architectures** όπου κάθε εργασία αντιστοιχίζεται στον βέλτιστο επεξεργαστή.

- Το **Flash-Lite** προτιμάται για σενάρια όπου η ταχύτητα και ο όγκος (throughput) είναι η βασική απαίτηση, και η πυκνότητα των οδηγιών είναι μέτρια (π.χ., γρήγορη ταξινόμηση δεδομένων, απλές απαντήσεις).

- Το **GPT-5 Nano** προτιμάται όταν η αξιοπιστία και η διατήρηση της ακρίβειας σε πιο σύνθετες, multi-step εργασίες (heavy rule stack) είναι απαραίτητες, ακόμη και με κόστος μεγαλύτερης καθυστέρησης.

Η παρατηρούμενη πτώση της ακρίβειας και των δύο μοντέλων σε πολύ υψηλό φόρτο οδηγιών επιβεβαιώνει ότι αυτά τα budget μοντέλα δεν μπορούν να υποκαταστήσουν τα Pro/Ultra μοντέλα σε σύνθετη, κρίσιμη συλλογιστική. Ως εκ τούτου, η χρήση τους πρέπει να περιορίζεται σε σαφώς καθορισμένες, επαναλαμβανόμενες εργασίες με αυστηρούς ελέγχους ποιότητας, καθώς ο κίνδυνος ανακριβών αποτελεσμάτων είναι υψηλός.[21]

Κεφάλαιο 8: Συμπεράσματα και Τελικές Συστάσεις

8.1 Σύνοψη των Τεχνικών και Επιχειρηματικών Επιλογών

Η τεχνολογική ανάπτυξη της Google, με επίκεντρο το Gemini Nano, αποτελεί μια αρχιτεκτονική απάντηση στις ελλείψεις των cloud-only μοντέλων, εστιάζοντας στο απόρρητο, την ανεξαρτησία από το δίκτυο και την εξάλειψη του κόστους token.[2] Το Gemini Nano είναι μια λύση ενσωματωμένη στο λειτουργικό σύστημα (AICore/Tensor chip) [1], η οποία επιτρέπει την ευρεία υιοθέτηση on-device AI μέσω εύχρηστων APIs (ML Kit).[9]

Αντίθετα, το οικοσύστημα ChatGPT/GPT (GPT-4/5) αντιπροσωπεύει τη λύση Cloud-First, η οποία υπερτερεί σε κλίμακα, στη σύνθετη συλλογιστική ικανότητα και στη διατήρηση υψηλής ακρίβειας σε απαιτητικές εργασίες.[4]

Η λειτουργία **Nano Banana** (Gemini 2.5 Flash Image) είναι η απάντηση της Google στον τομέα της δημιουργικής πολυτροπικότητας, με ιδιαίτερη έμφαση στη διατήρηση της συνοχής κατά την επαναληπτική επεξεργασία εικόνας.[4]

8.2 Προτάσεις για Υλοποίηση Βάσει Προτεραιοτήτων (Recommendations)

Για τους τεχνικούς υπεύθυνους προϊόντων και τους προγραμματιστές, η επιλογή του

κατάλληλου μοντέλου πρέπει να βασίζεται σε ένα δέντρο αποφάσεων κόστους-ρίσκου, με βάση τον επιχειρησιακό στόχο:

Προτεραιότητα / Στόχος	Συνιστώμενο Μοντέλο	Λόγος Σύστασης
Απόρρητο & Offline Λειτουργία	Gemini Nano (On-Device / ML Kit)	Τα δεδομένα επεξεργάζονται τοπικά, εξάλειψη δικτυακής καθυστέρησης, μηδενικό κόστος κλήσης.[2]
Υψηλότερη Γνωστική Ικανότητα / Κώδικας	GPT-5 / Gemini Pro (Cloud)	Υψηλή ακρίβεια (π.χ., 88.4% GPQA) και ικανότητα σύνθετης συλλογιστικής και ανάλυσης μεγάλων context windows.[4]
Γρήγορος Όγκος Δεδομένων (Low Load Cloud)	Gemini 2.5 Flash-Lite (Cloud)	Ηγέτης στην ταχύτητα, βέλτιστη λύση κόστους-ταχύτητας για απλές εργασίες ταξινόμησης και throughput.[21]
Δημιουργία & Επεξεργασία Εικόνας με Συνοχή	Nano Banana / Gemini 2.5 Flash Image	Εξειδίκευση στη συνεπή επεξεργασία χαρακτήρων και διατήρηση σκηνής σε επαναληπτικές ροές εργασίας.[3]
Εξοικονόμηση Κόστους Εισόδου (Input Cost)	GPT-5 Nano (Cloud)	Φθηνότερο κόστος εισόδου token ($0.05/1M tokens) για εργασίες με βαριά είσοδο αλλά ελαφριά έξοδο.[21]

8.3 Μελλοντικές Τάσεις στα On-Device LLMs

Η εξέλιξη των on-device LLMs υποδεικνύει τρεις κύριες μελλοντικές τάσεις:

1. **Ενισχυμένη Κβαντοποίηση (Quantization) και Συμπίεση:** Η συνεχής έρευνα για τεχνικές όπως η κβαντοποίηση (μείωση των ψηφίων από 32-bit σε 8-bit ή 4-bit) επιτρέπει τη δραστική μείωση του μεγέθους των μοντέλων και της απαιτούμενης μνήμης RAM, διατηρώντας την ακρίβεια. Αυτό θα επιτρέψει σε ακόμη μεγαλύτερα μοντέλα να εκτελούνται στη συσκευή, μειώνοντας περαιτέρω το χάσμα ικανότητας με τα cloud LLMs.
2. **Πλήρης Πολυτροπική Ολοκλήρωση On-Device:** Το Nano 2.0 με την υποστήριξη εικόνας [5] σηματοδοτεί την έναρξη μιας εποχής όπου οι συσκευές θα μπορούν να επεξεργάζονται εικόνες, ήχο και κείμενο ταυτόχρονα, εξ ολοκλήρου στη συσκευή. Αυτή η ικανότητα είναι κρίσιμη για προηγμένες λειτουργίες ασφαλείας και υγείας, όπως η βελτιωμένη
   Scam Detection και η επεξεργασία δεδομένων υγείας (*health data processing*), οι οποίες ήδη χρησιμοποιούν την προστασία του Tensor/Titan M2.
3. **Επέκταση Πλατφόρμας:** Η διάθεση του **LiteRT-LM** [2], του μηχανισμού συμπερασμάτων που τροφοδοτεί το Nano, υποδεικνύει τη στρατηγική της Google να επεκτείνει τις on-device δυνατότητες πέρα από το Android, σε άλλα περιβάλλοντα όπως ο Chrome, το Chromebook Plus και το Pixel Watch. Αυτό εξασφαλίζει τη μεγαλύτερη δυνατή εμβέλεια και τη διατήρηση της αρχιτεκτονικής συνέπειας για την on-device AI.