Google Duplex in Gemini explained: how an AI agent phones businesses for you

22 November 2026 · 6 min read · blogs.iictp.org/google-duplex-gemini-tasks-phone-calls-explained

Short answer: Google Duplex is the conversational calling system Google first introduced in 2018, and Google Research says it now powers phone calls to businesses inside Tasks in the Gemini app, launched in the week of its 20 November 2026 research post. According to the research post of 20 November, the stack moved from narrow custom neural models to a single design in which Gemini handles understanding, planning and dialogue, which matters technically because a phone call is a live, interruptible, multi-turn task rather than a text prompt.

TL;DR The short version

- Google Duplex, introduced in 2018, now powers phone calls to businesses from Tasks in the Gemini app, according to Google Research.
- The stack has moved from narrow custom models to a design in which Gemini handles understanding, planning and dialogue.
- A live phone call is a hard agent problem because it is fast and interruptible; the post is Google's own description, with no independent evaluation cited.

Read online, listen to it, or change the text size: https://blogs.iictp.org/google-duplex-gemini-tasks-phone-calls-explained

What is Google Duplex?

Duplex is a system that holds natural voice conversations with people at local businesses, for example to book a table or an appointment. Google Research authors Yossi Matias, Yaniv Leviathan and Janet Zhou describe it as shared conversational infrastructure that has been used across several Google products, and now as the engine behind calling businesses on a user’s behalf in Tasks in Gemini.

How it works: three eras of the stack

- **Era 1, narrow domains.** Before modern language models, Duplex used custom recurrent neural networks trained on anonymised, task-specific phone conversations. Without pre-trained world knowledge, each scenario had to be taught explicitly.
- **Era 2, hybrid stacks.** As open-domain foundation models arrived, Duplex used them in production; the post says LaMDA was used in live calls in 2022. Because those models were less predictable and sometimes slower, the team paired generative models for flexible phrasing with deterministic state tracking and strict policy guardrails.

- **Era 3, task prompts and Gemini.** The fixed verticals were dropped. A task is described in free text, and Gemini unifies multimodal understanding, planning and dialogue so the agent can handle multi-constraint requests, such as booking car maintenance around calendar conflicts, and adapt when a receptionist offers a different slot.

Why a phone call is a hard agent problem

The post explains the difficulty in its own words: live calls move fast, the other person can revise a thought mid-sentence, and turn-taking depends on subtle cues, so a pause that is half a second too long sounds like a dropped call. A text agent can take seconds to think; a calling agent cannot. That is why the hybrid pattern of flexible generation plus deterministic state and policy checks is worth understanding even if you never build voice agents.

What the source reports

- Business information sourced from Duplex calls had been shown in Search and Maps over one trillion times, a figure Google says it shared in 2023.
- Conversational models were brought on-device from 2019 through Pixel Call Assist across five consecutive Pixel generations, powering features such as Call Screen and Hold for Me.
- During the COVID-19 pandemic Duplex verified merchant information automatically to keep listings accurate.

Limitations and open questions

- The post is a Google description of its own system. It was read only in part; success rates, latency, languages, supported countries and how businesses are told they are speaking to an AI were not read and are not claimed here.
- No independent evaluation of call outcomes is cited in the post.
- Availability may differ by country and account type; check Google’s product pages.

What it means for builders of agents

If you build an agent that acts in the real world, the useful lesson is architectural. Keep the model for language and judgement, and keep state, permissions and policy in deterministic code you can test. See AI agents explained and how to evaluate an LLM application.

Frequently asked questions

- **Is Google Duplex new?** No. It was introduced in 2018. What is new, per Google’s 20 November 2026 post, is its launch inside Tasks in the Gemini app.
- **Does it use Gemini?** The post says the calling stack uses Gemini to unify multimodal understanding, planning and dialogue.
- **Is this peer reviewed research?** No. It is a company research blog post.

Related reading

- AI agents explained
- What is prompt injection?

- Responsible AI in practice
- IICTP's AI Practitioner credential covers agent design and evaluation.

Sources and currency

Sources were read on 22 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Yossi Matias, Yaniv Leviathan, Janet Zhou (Google Research). How Google Duplex evolved to also call businesses for you in Gemini. 20 November 2026. research.google/blog/how-google-duplex-evolved-to-also-call-businesses-for-you-in-gemini.

Published by IICTP, an independent technical credential body, on 22 November 2026. The current version of this post is at https://blogs.iictp.org/google-duplex-gemini-tasks-phone-calls-explained. More guides: blogs.iictp.org.