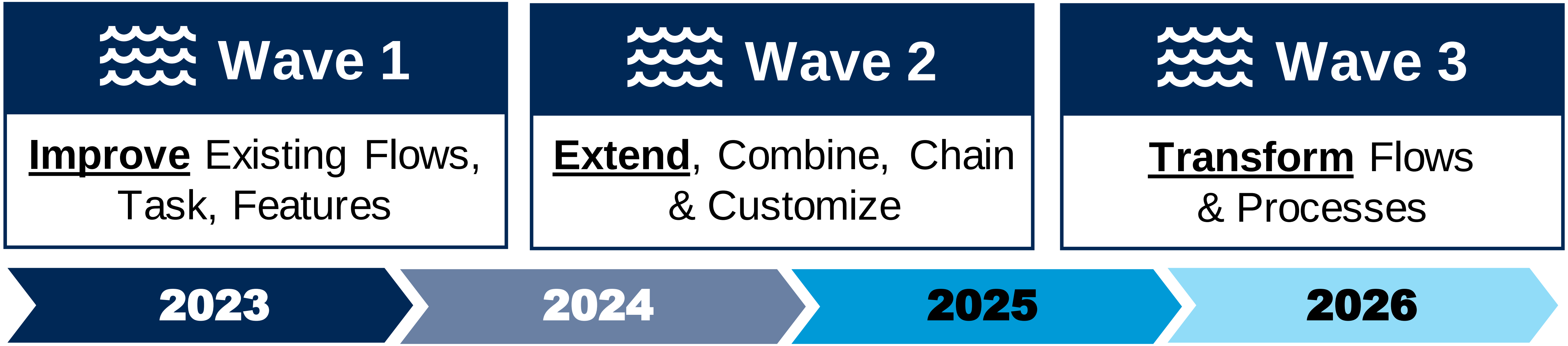
Generative AI in Customer Service: Deployments, Pilots and Best Practices

Bern Elliot

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's research organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner research may address legal and financial issues, Gartner does not provide legal or investment advice and its research should not be construed or used as such. Your access and use of this publication are governed by Gartner's Usage Policy. Gartner prides itself on its reputation for independence and objectivity. Its research is produced independently by its research organization without input or influence from any third party. For further information, see "Guiding Principles on Independence and Objectivity."

Maturing LLM Use in Customer Service/Support



Wave 1

Improve Existing Flows, Task, Features

Wave 2

Extend, Combine, Chain & Customize

Wave 3

Transform Flows & Processes

Tech: Out-of-Box LLM & Prompt Engineering

- Limited technical skills
- Rarely customer facing
- Rapid improvements
- Emerging governance

Tech: Fine-Tuning, Retraining, LLM Chains

- Advanced tech skills
- Multiple models & data
- Ongoing expertise
- Centralizing governance

Tech: Orchestration Engine, End-to-End Flows

- Much still in “lab” phase
- Intentless dialog flows
- Design virtual assistants
- Integrated governance

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Enabling Governance, Risk Mitigation & Controls

Configuration	Granular access control
Filtering	PII, e-PHI, sensitive data, threat/violence, hallucination, profanity
Metering	Cost limits, latencies, availability
Audit Log	Reports on filtering & metering

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

More Details on Filtering Governance

avaamo

Filtering

The following PII and sensitive data entities will be substituted with the equivalent values prior to a call to the GPT model

Personal Identifiable Information (PII)

These PII fields in either the request or response will be masked, substituted by the platform.

Name	**Date of Birth**	**Email**	**Address**

Sensitive Data

These sensitive data fields in either the request or response will be substituted automatically by the platform.

Transaction IDs	**Date/Time**	**Amount**	**Location**

Electronic Personal Health Information (e-PHI)

These e-PHI fields in either the request or response will be masked, substituted by the platform.

Medical Record	**Lab Report**	**Appointments**	**Procedures**

Key Filtering Features:

- Identification
- Replacement
- Audit
- Customization

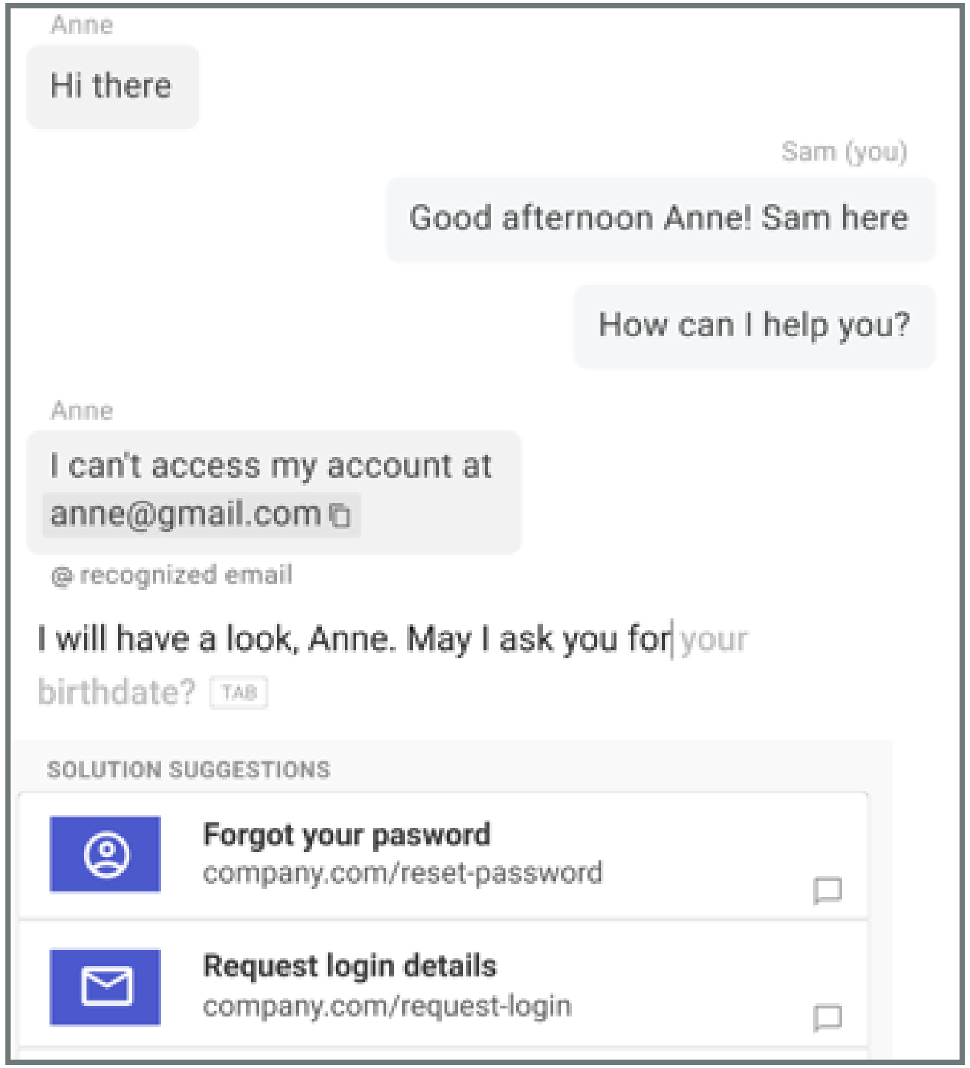
Example Deployments:

ERICSSON intel pepsi

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Fine Tuning LLMs for Agent Assist

deepdesk



Anne

Hi there

Sam (you)

Good afternoon Anne! Sam here

How can I help you?

Anne

I can't access my account at anne@gmail.com

@ recognized email

I will have a look, Anne. May I ask you for your birthdate? TAB

SOLUTION SUGGESTIONS

Forgot your pasword
company.com/reset-password

Request login details
company.com/request-login

Sentence autocomplete for live agent chat and messaging agent assist.

- 20% fine-tuned GPT2, rest from their own models (some LLM, some other).
- Customize text "tone" per agent style.
- On average agents don't type 50% of their responses.

Fine-tuning GPT2:

Data from three months, ~250K-1M conversations. Cost is ~3K euros/month.

If adding GPT3/4, issues include scaling, cost, governance, confidentiality. BYO GPT3/4 (~$1M/year for 500K conversations/month).

Using These Methods: DHL Rabobank vodafone

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

GenAI – Enabling Applications

OneReach.ai

Fortune 50 GenAI assistant. In limited production, expanding to 300K global users.
Usage areas include customer service, HR, SCM

Solution: Concierge digital workers integrate 4 skills with applications

- Summarization
- Writing assistant
- Simplification
- Document analysis

Decision Drivers:

- Not locked into a single enterprise app vendor
- Able to customize for use with channel and application
- Control over governance
- Cost and consumption controlled on per-user level

Application Approach:

- Use transparent, deterministic, auditable models to develop answers initially
- Use LLM for creating smooth text from the answer — “guardrails”
- Allow experimentation with LLM for generating answers, but this can be turned off

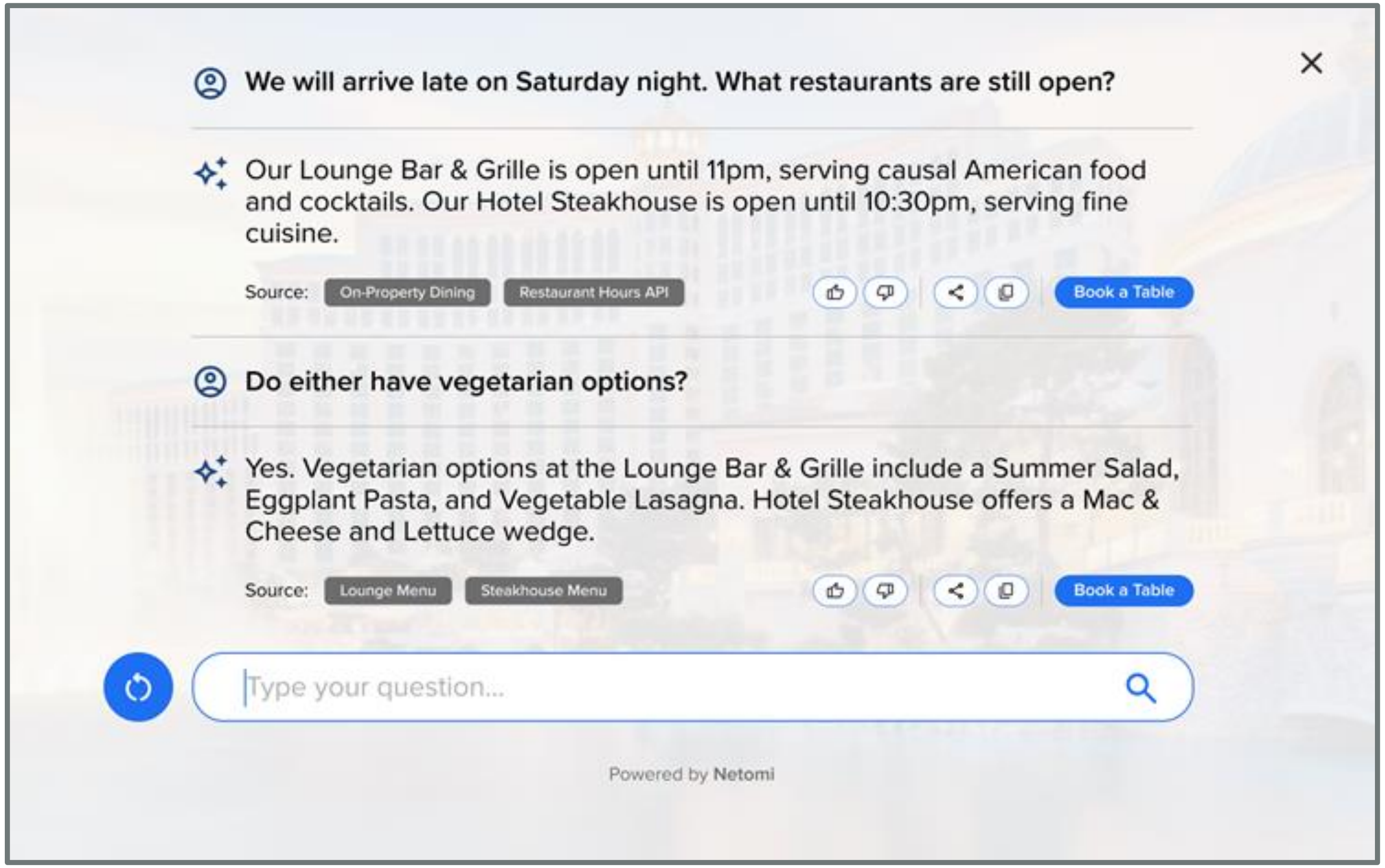
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Conversational Search with Generative AI

Conversational responses from multiple sources, carries context forward through interactions

Phases of request handling:

- **Transform & enrich Messages**
  Prepare input for information retrieval and task execution
- **Execution planning**
  Identify tasks, models and prepare prompts with Netomi's SanctionedAI™
- **Data retrieval & composition** Combine knowledge and data from APIs, validate with Netomi's SanctionedAI™ for brand safety
- **Generate response** with source reference for transparency



We will arrive late on Saturday night. What restaurants are still open?

Our Lounge Bar & Grille is open until 11pm, serving causal American food and cocktails. Our Hotel Steakhouse is open until 10:30pm, serving fine cuisine.

Source: On-Property Dining | Restaurant Hours API

Do either have vegetarian options?

Yes. Vegetarian options at the Lounge Bar & Grille include a Summer Salad, Eggplant Pasta, and Vegetable Lasagna. Hotel Steakhouse offers a Mac & Cheese and Lettuce wedge.

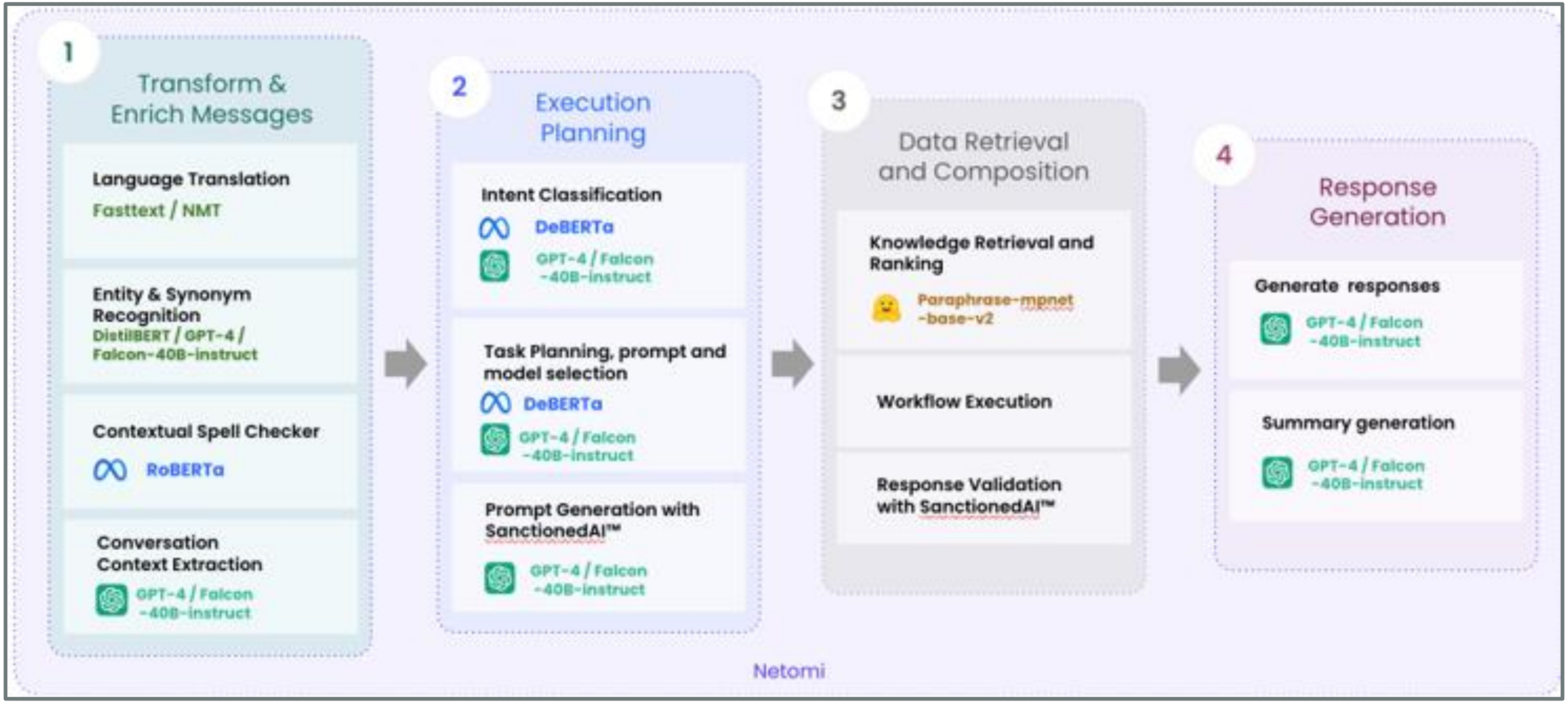
Source: Lounge Menu | Steakhouse Menu

Deployed at a Leading Entertainment and Resort Company

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

LLM Deployment Optimized for Predictable, Transparent and Brand Safe Execution

netomi



1 Transform & Enrich Messages

Language Translation
Fasttext / NMT

Entity & Synonym Recognition
DistilBERT / GPT-4 / Falcon-40B-instruct

Contextual Spell Checker
RoBERTa

Conversation Context Extraction
GPT-4 / Falcon -40B-instruct

2 Execution Planning

Intent Classification
DeBERTa
GPT-4 / Falcon -40B-instruct

Task Planning, prompt and model selection
DeBERTa
GPT-4 / Falcon -40B-instruct

Prompt Generation with SanctionedAI™
GPT-4 / Falcon -40B-instruct

3 Data Retrieval and Composition

Knowledge Retrieval and Ranking
Paraphrase-mpnet -base-v2

Workflow Execution

Response Validation with SanctionedAI™

4 Response Generation

Generate responses
GPT-4 / Falcon -40B-instruct

Summary generation
GPT-4 / Falcon -40B-instruct

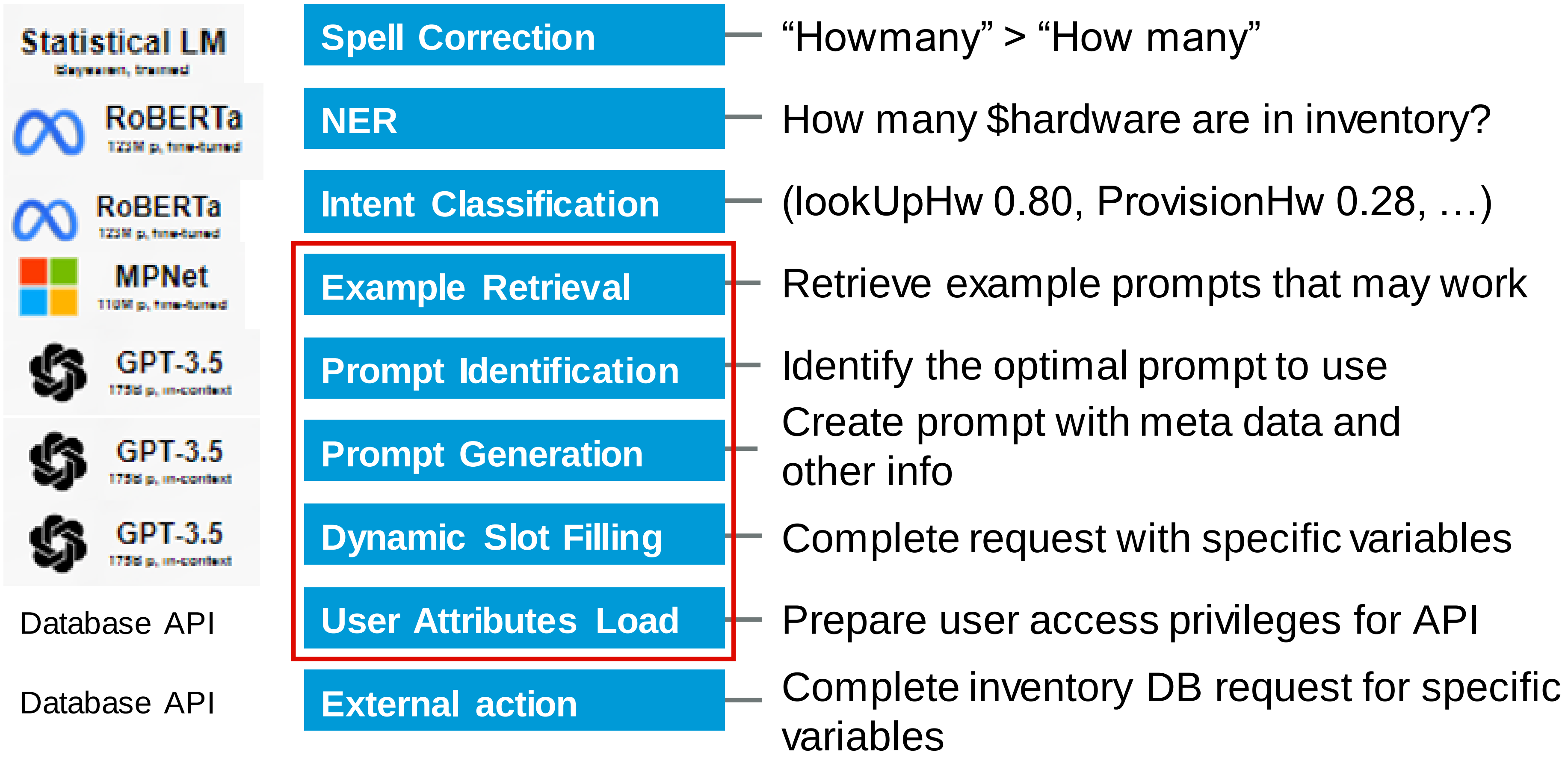
Netomi

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

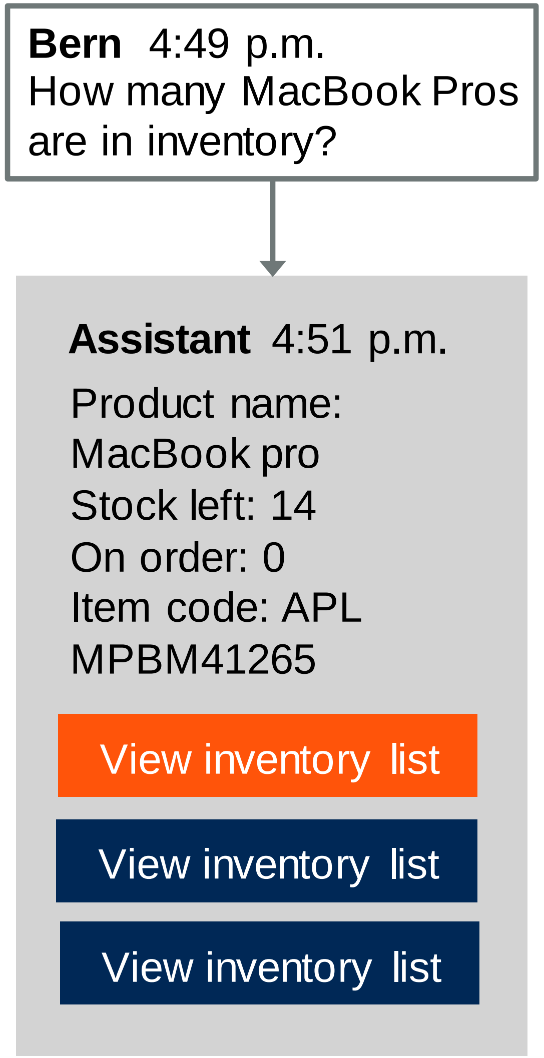
Chaining LLMs for an End-to-End Task

Moveworks

Chain Different Models for Different Purposes



Statistical LM Bayesian, trained	Spell Correction	“Howmany” > “How many”
RoBERTa 123M p, fine-tuned	NER	How many $hardware are in inventory?
RoBERTa 123M p, fine-tuned	Intent Classification	(lookUpHw 0.80, ProvisionHw 0.28, …)
MPNet 110M p, fine-tuned	Example Retrieval	Retrieve example prompts that may work
GPT-3.5 175B p, in-context	Prompt Identification	Identify the optimal prompt to use
GPT-3.5 175B p, in-context	Prompt Generation	Create prompt with meta data and other info
GPT-3.5 175B p, in-context	Dynamic Slot Filling	Complete request with specific variables
Database API	User Attributes Load	Prepare user access privileges for API
Database API	External action	Complete inventory DB request for specific variables



Bern 4:49 p.m.
How many MacBook Pros are in inventory?

Assistant 4:51 p.m.
Product name: MacBook pro
Stock left: 14
On order: 0
Item code: APL MPBM41265

250+ Deployments, Including: ALBEMARLE Seagen TOYOTA

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Gartner

Findings and Recommendations

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Maturing LLM Use in Customer Service/Support

Wave 1

Improve Existing Flows, Task, Features

Wave 2

Extend, Combine, Chain & Customize

Wave 3

Transform Flows & Processes

2023
2024
2025
2026

Tech: Out-of-Box LLM & Prompt Engineering

- Summarization, simplification, intent handling, multimodel AI
- Emerging governance

Tech: Fine-Tuning, Retraining, LLM Chains

- Retrieval augmented generation, chaining, fine-tuning (simple)
- Centralized governance

Tech: Orchestration Engine, End-to-End Flows

- Not in production
- Intentless dialog flows
- Design virtual assistants
- Integrated governance

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

LLMs in Customer Service Trends

Business	Market hype creating unrealistic user expectations. Concerns regarding price of large models.
Organization	GenAI governance planning is a critical first step. Employee-facing is by far the most common use case.
Technology	What works in POC may not work in production. Solutions combine GenAI and other AI methods. Interest growing in fine-tuning proprietary LLMs.

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Planning Your LLM Customer Service Roadmap

User Profile	Now	Two Years From Now
Modest Adoption	Case studies & pilots to understand LLM uses and vendor partner plans.	Incremental expansion working with vendor solution. Advanced data integrations
Advanced Adoption	Establish vision and roadmap. Use preintegrated solutions. Identify unique data needs.	Add GenAI to AI skills. Use data to advance model tuning and prompt usage.
Aggressive Adoption	Be part of your organization’s broader generative AI strategy. Identify your data & models.	Redesigned CS process flows for increase augmentation & automating.

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Recommended Gartner Research

- **Tool: Enterprise Use Cases for ChatGPT**
  Anthony Mullen, Wilco van Ginkel and Others
- **Use-Case Prism: Artificial Intelligence for Customer Service**
  Bern Elliot and Wynn White
- **How Can Generative AI Be Used to Improve Customer Service and Support?**
  Pri Rathnayake
- **How to Pilot Generative AI**
  Leinar Ramos, Anthony Mullen and Others
- **AI Design Patterns for Large Language Models**
  Leinar Ramos, Anthony Mullen and Others
- **Applying AI — A Framework for the Enterprise**
  Bern Elliot, Anthony Mullen and Erick Brethenoux

Access to Gartner research is subject to entitlement. For information, please contact your Gartner representative.

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Thank You

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.