blp connect

Custom Agents Deep Dive

Was Custom Agents sind und wie wir sie bei BLP einsetzen

Claudio Ferrari

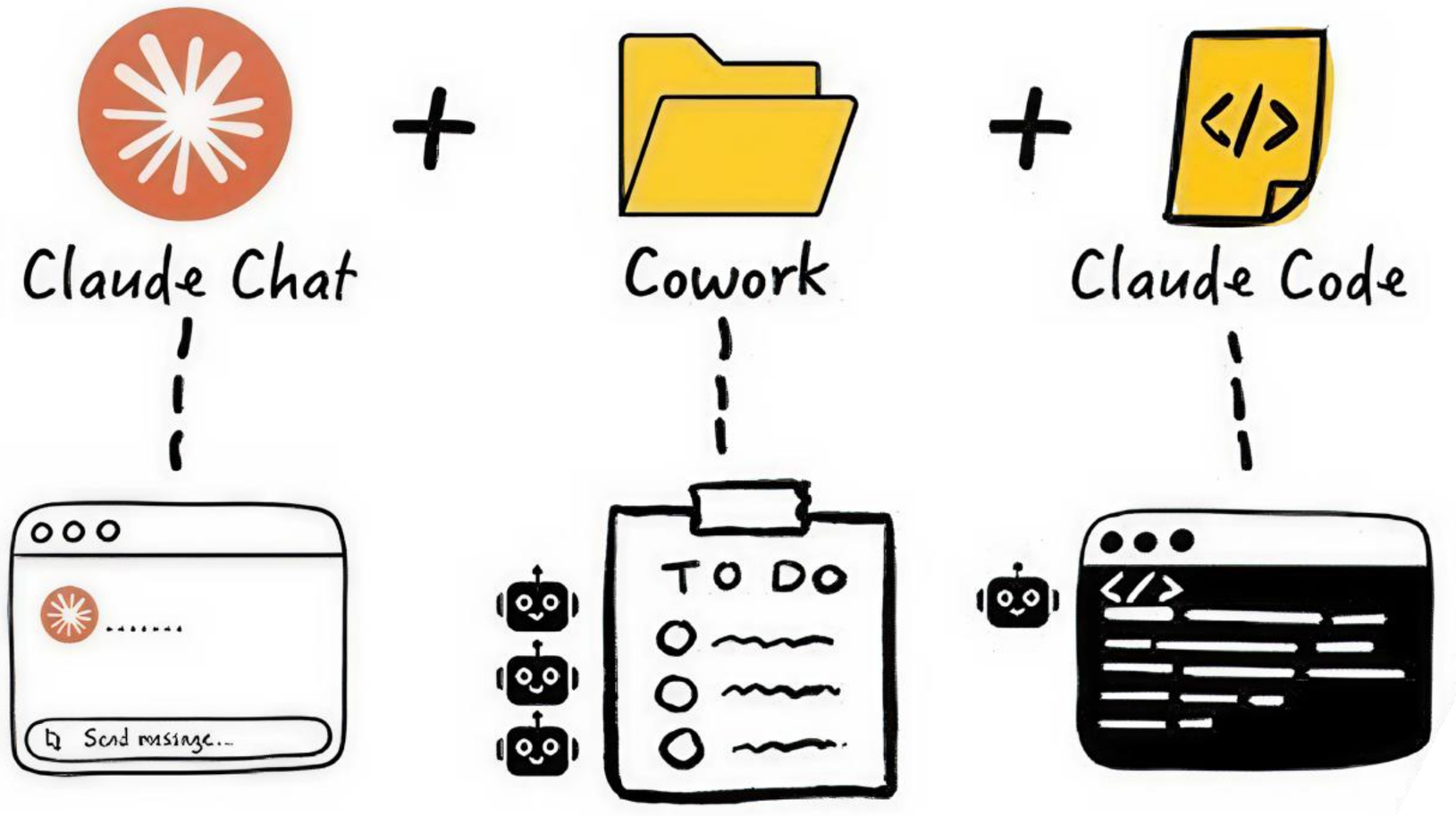
Technical Lead, AI Agents

blp

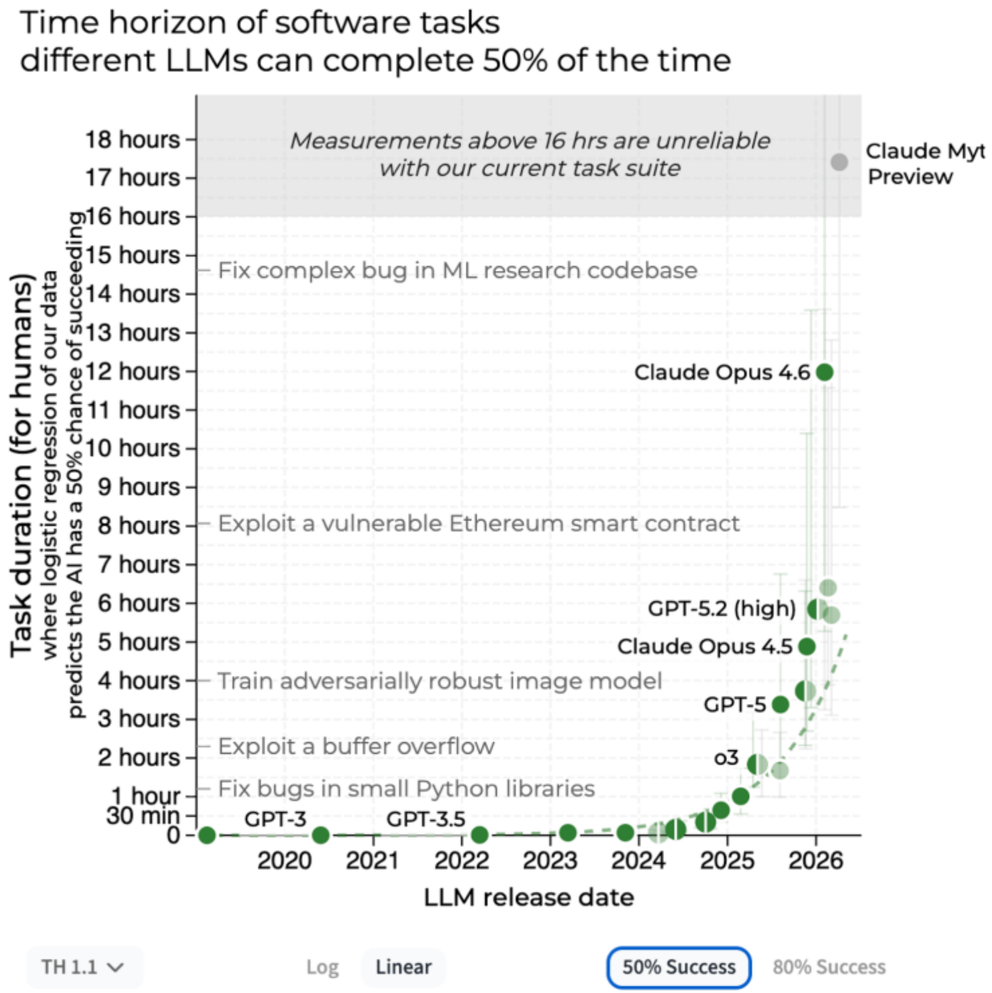

ChatGPT


Gemini


Claude


Claude Chat
Cowork
Claude Code
Send message...
TO DO

KI-Modelle werden immer besser



KI-Halluzinationen

Trainings- und Evaluierungsverfahren belohnen

das Raten stärker als das Eingestehen von Unsicherheit.

Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI
Ofir Nachum OpenAI

September

Abstract

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

Leaderboard

We update questions regularly so that the benchmark completely refreshes every 6 months. Some questions for previous releases are available here. The most recent version is **LiveBench-2026-01-08**. This version features a new mathematical task and a new data analysis task.

To further reduce contamination, we delay publicly releasing the questions from the most-recent updates.

View Full Changelog

2026-01-08

Agentic Coding Average | Mathematics Average | Data Analysis Average | Language Average | IF Average

Show Subcategories | Show Open Weight Models Only | Show Model Effort Variants | Show High Unseen Question Bias Models | Clear Filters

Filter by organization...

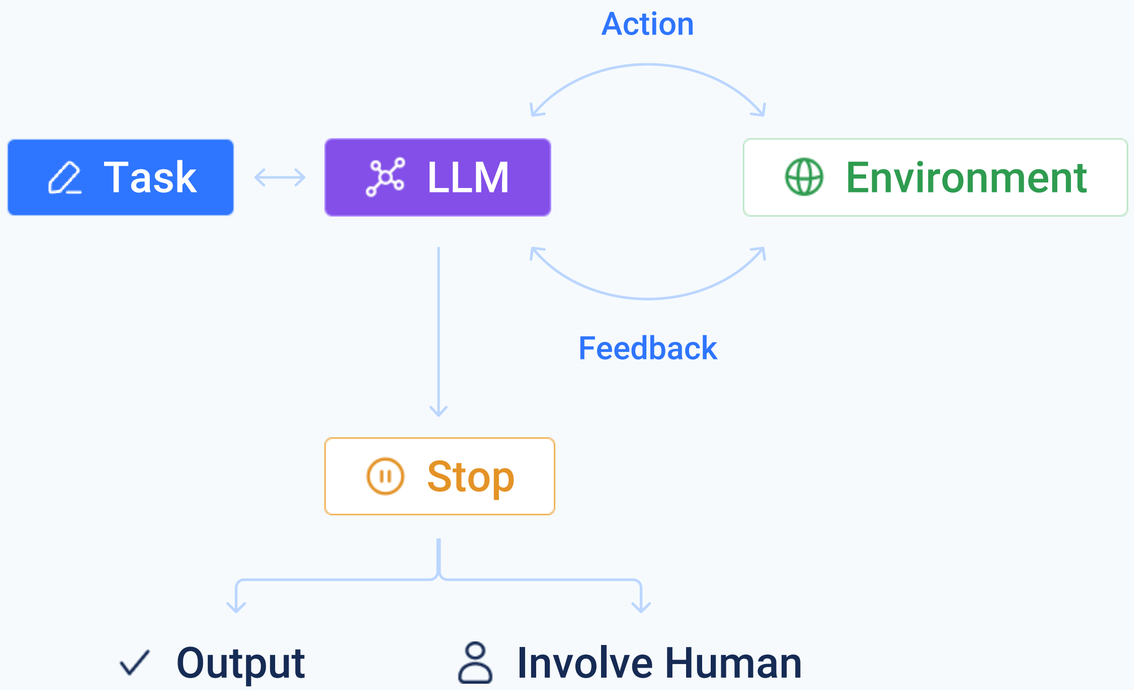
Model	Organization	Global Average	Reasoning Average	Coding Average	Agentic Coding Average	Mathematics Average	Data Analysis Average	Language Average	IF Average
GPT-5.5 Thinking xHigh Effort	OpenAI	80.71	87.71	82.47	56.67	96.32	81.08	87.66	73.04
GPT-5.4 Thinking xHigh Effort	OpenAI	80.28	88.12	77.54	70.00	94.15	79.31	82.63	70.22
Gemini 3.1 Pro Preview High* *5th rank in unseen questions across all categories	Google	79.93	84.00	76.45	65.00	91.04	78.54	85.38	79.10
Claude 4.7 Opus Thinking xHigh Effort	Anthropic	76.91	87.69	82.09	60.00	93.10	78.26	77.91	59.34

blp connect

Was ist ein Agent?

01

KI-Agenten



Tools: Führen Aufgaben aus

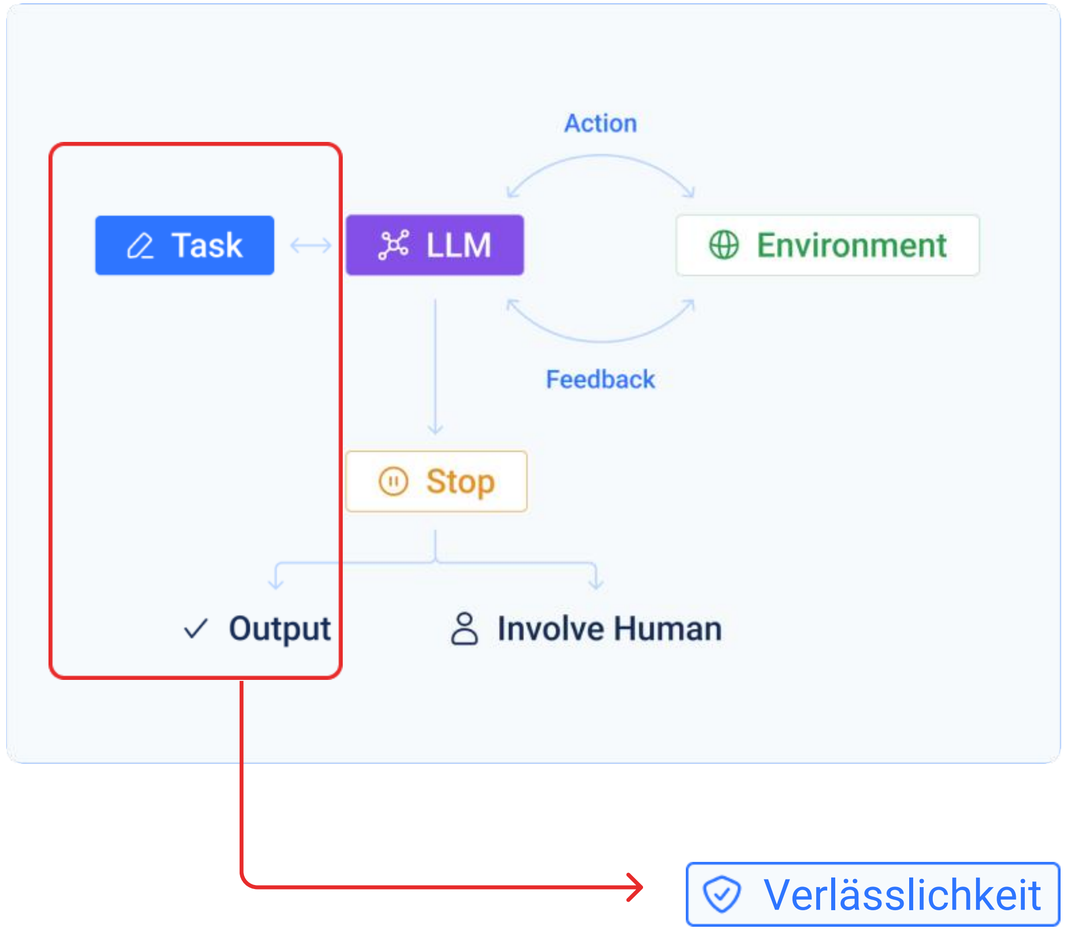
Kontext: E-Mails, Dokumente,...

MCP: ERP- und Workflow-Daten...

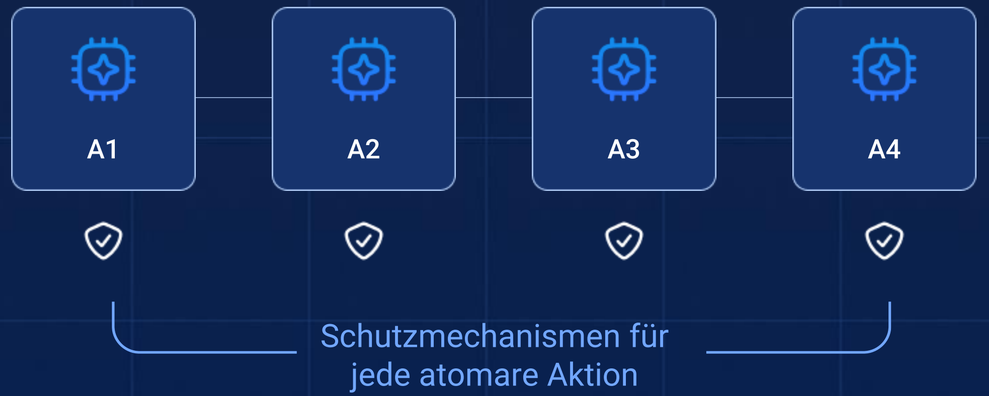
Standard-Verfahren:
Unternehmensrichtlinien

Erinnerungs-Komponente:
Behält informationen aus früheren Interaktionen und Nutzeranweisungen

BLPs Atomare Agenten

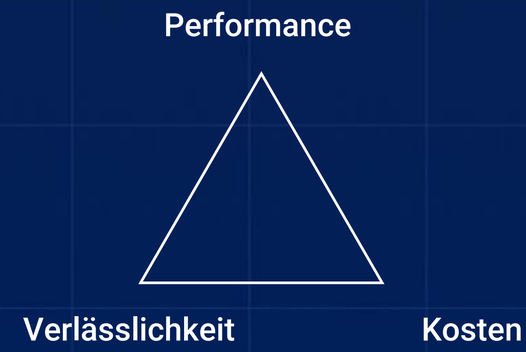


Workflow verteilt auf spezialisierte Agenten



Das bedeutet atomar bei BLP

- aufgesplittet in Teilaufgaben
- Unabhängig testbar & austauschbar
- dort abgesichert, wo es sinnvoll ist



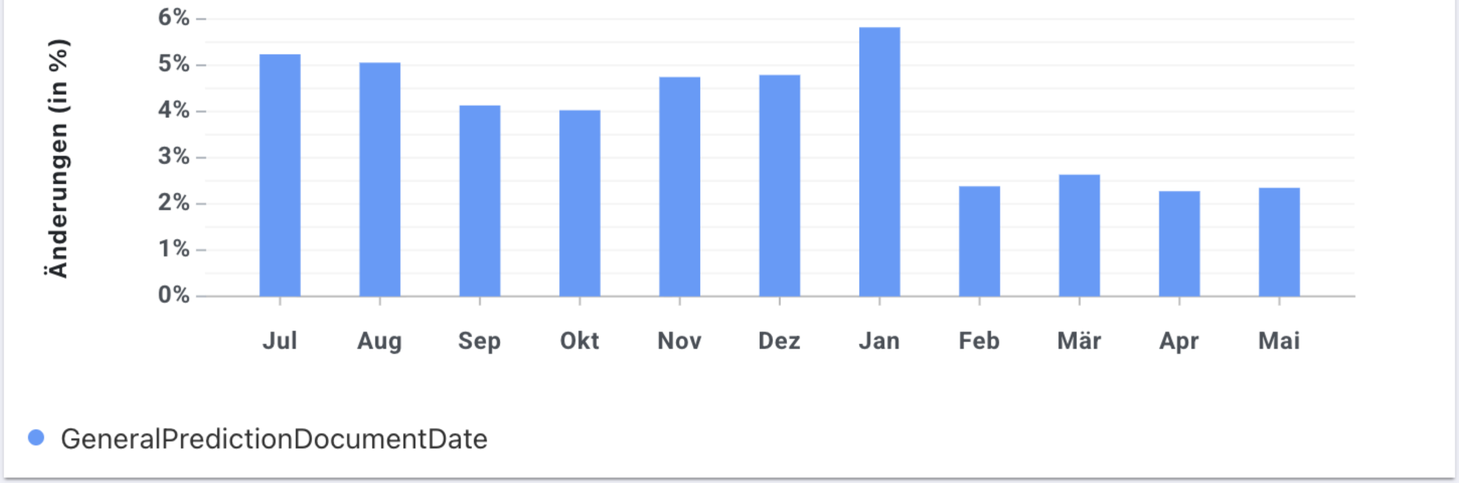
blp connect

Wofür setzen wir Agenten ein?

1. Standard
2. Custom AI
3. Meta-Agenten
4. Standard-Erweiterungen
5. Unbound

02

Document Date Changes per month (Stand: 11.05.2026, 09:05:01)

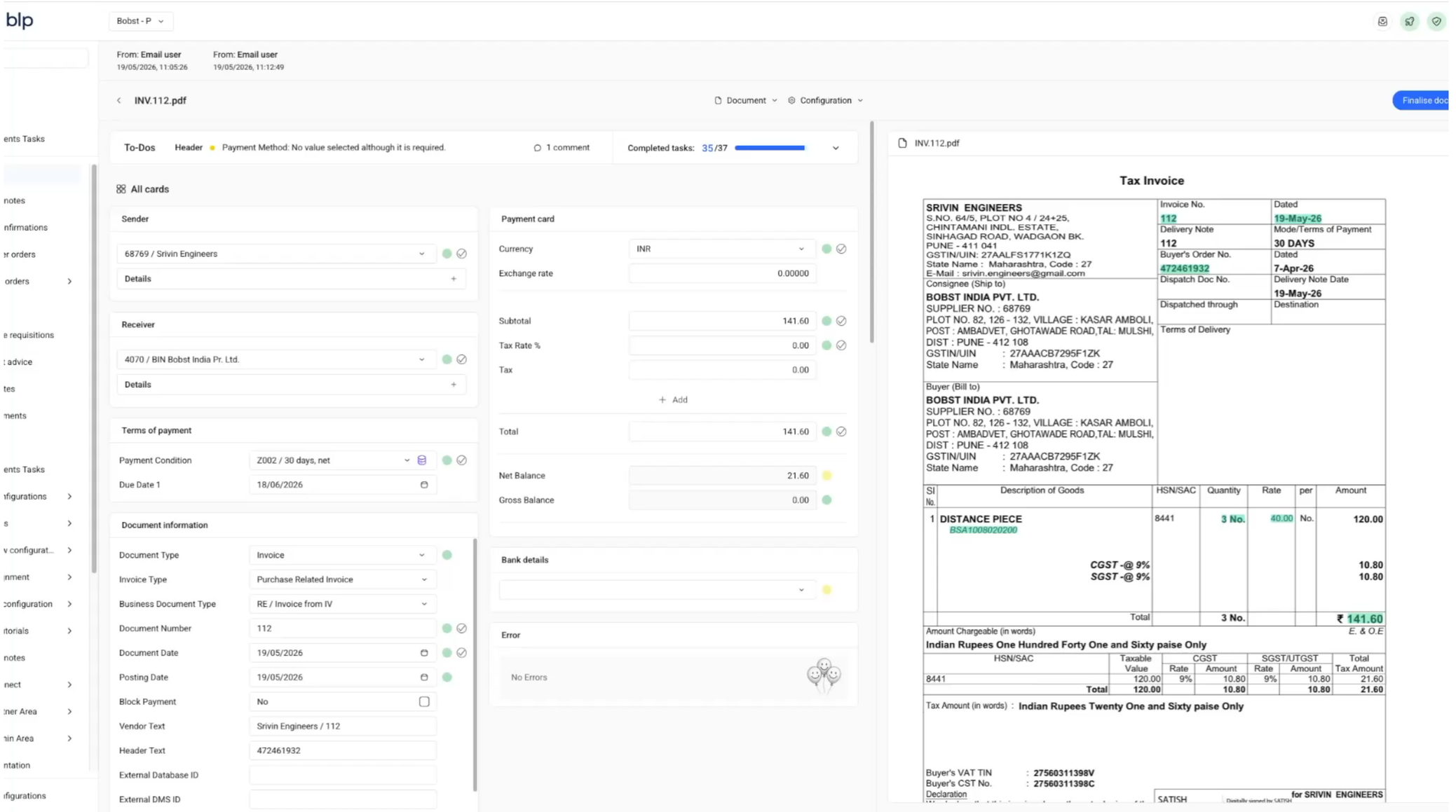


blp connect

Wofür setzen wir Agenten ein?

1. Standard
2. Custom AI
3. Meta-Agenten
4. Standard-Erweiterungen
5. Unbound

02


blp
Bobst - P
From: Email user
19/05/2026, 11:05:26
From: Email user
19/05/2026, 11:12:49
INV.112.pdf
Document
Configuration
Finalise doc
To-Dos
Header
Payment Method: No value selected although it is required.
1 comment
Completed tasks: 35/37
All cards
Sender
68769 / Srivin Engineers
Details
Receiver
4070 / BIN Bobst India Pr. Ltd.
Details
Terms of payment
Payment Condition
Z002 / 30 days, net
Due Date 1
18/06/2026
Document information
Document Type
Invoice
Invoice Type
Purchase Related Invoice
Business Document Type
RE / Invoice from IV
Document Number
112
Document Date
19/05/2026
Posting Date
19/05/2026
Block Payment
No
Vendor Text
Srivin Engineers / 112
Header Text
472461932
External Database ID
External DMS ID
Payment card
Currency
INR
Exchange rate
0.00000
Subtotal
141.60
Tax Rate %
0.00
Tax
0.00
Add
Total
141.60
Net Balance
21.60
Gross Balance
0.00
Bank details
Error
No Errors
INV.112.pdf
Tax Invoice
SRIVIN ENGINEERS
S.NO. 64/5, PLOT NO 4 / 24+25,
CHINTAMANI INDL. ESTATE,
SINHAGAD ROAD, WADGAON BK.
PUNE - 411 041
GSTIN/UIN: 27AALFS1771K1ZQ
State Name : Maharashtra, Code : 27
E-Mail : srivin.engineers@gmail.com
Invoice No.
112
Dated
19-May-26
Delivery Note
112
Mode/Terms of Payment
30 DAYS
Buyer's Order No.
472461932
Dated
7-Apr-26
Dispatch Doc No.
Delivery Note Date
19-May-26
Dispatched through
Destination
Terms of Delivery
Consignee (Ship to)
BOBST INDIA PVT. LTD.
SUPPLIER NO. : 68769
PLOT NO. 82, 126 - 132, VILLAGE : KASAR AMBOLI,
POST : AMBADVET, GHOTAWADE ROAD,TAL: MULSHI,
DIST : PUNE - 412 108
GSTIN/UIN : 27AACB7295F1ZK
State Name : Maharashtra, Code : 27
Buyer (Bill to)
BOBST INDIA PVT. LTD.
SUPPLIER NO. : 68769
PLOT NO. 82, 126 - 132, VILLAGE : KASAR AMBOLI,
POST : AMBADVET, GHOTAWADE ROAD,TAL: MULSHI,
DIST : PUNE - 412 108
GSTIN/UIN : 27AACB7295F1ZK
State Name : Maharashtra, Code : 27
Sl No. | Description of Goods | HSN/SAC | Quantity | Rate | per | Amount
1 | DISTANCE PIECE BSA1008020200 | 8441 | 3 No. | 40.00 | No. | 120.00
CGST -@ 9% | 10.80
SGST -@ 9% | 10.80
Total | 3 No. | ₹ 141.60
Amount Chargeable (in words)
E. & O.E
Indian Rupees One Hundred Forty One and Sixty paise Only
HSN/SAC | Taxable Value | CGST Rate | CGST Amount | SGST/UTGST Rate | SGST/UTGST Amount | Total Tax Amount
8441 | 120.00 | 9% | 10.80 | 9% | 10.80 | 21.60
Total | 120.00 | | 10.80 | | 10.80 | 21.60
Tax Amount (in words) : Indian Rupees Twenty One and Sixty paise Only
Buyer's VAT TIN : 27560311398V
Buyer's CST No. : 27560311398C
Declaration
for SRIVIN ENGINEERS
SATISH

blp connect

Wofür setzen wir Agenten ein?

1. Standard
2. Custom AI
3. Meta-Agenten
4. Standard-Erweiterungen
5. Unbound

02

Agenten, die deine Agenten verbessern — während du schläfst.

TAG **Euer Team arbeitet**

Korrekturen und manuelle Überschreibungen von Agenten-Entscheidungen

Rückfragen und menschliche Antworten auf eskalierte Fälle

Analytics Layer: Jedes Signal wird zu Trainingsdaten auf dem Digital Twin.

NACHT **Atomare Agenten übernehmen. Der Auto-Improvement-Cycle läuft.**

01
Evals generieren
Echte Fälle werden zu Testdaten.

02
SOPs tunen
Prozesslogik schärfen — ohne Code-Änderung.

03
Validieren
Neue Agenten gegen das Eval-Set laufen lassen.

04
Ausrollen
Verbesserte Agenten live am Morgen. Reversibel.

Meta Agent Runs

Neuen Run erstellen

Sachkonto Meta Agent

Läuft... | Abgeschlossen | Fehlgeschlagen | Abgebrochen

Noch keine Meta Agent Runs vorhanden...

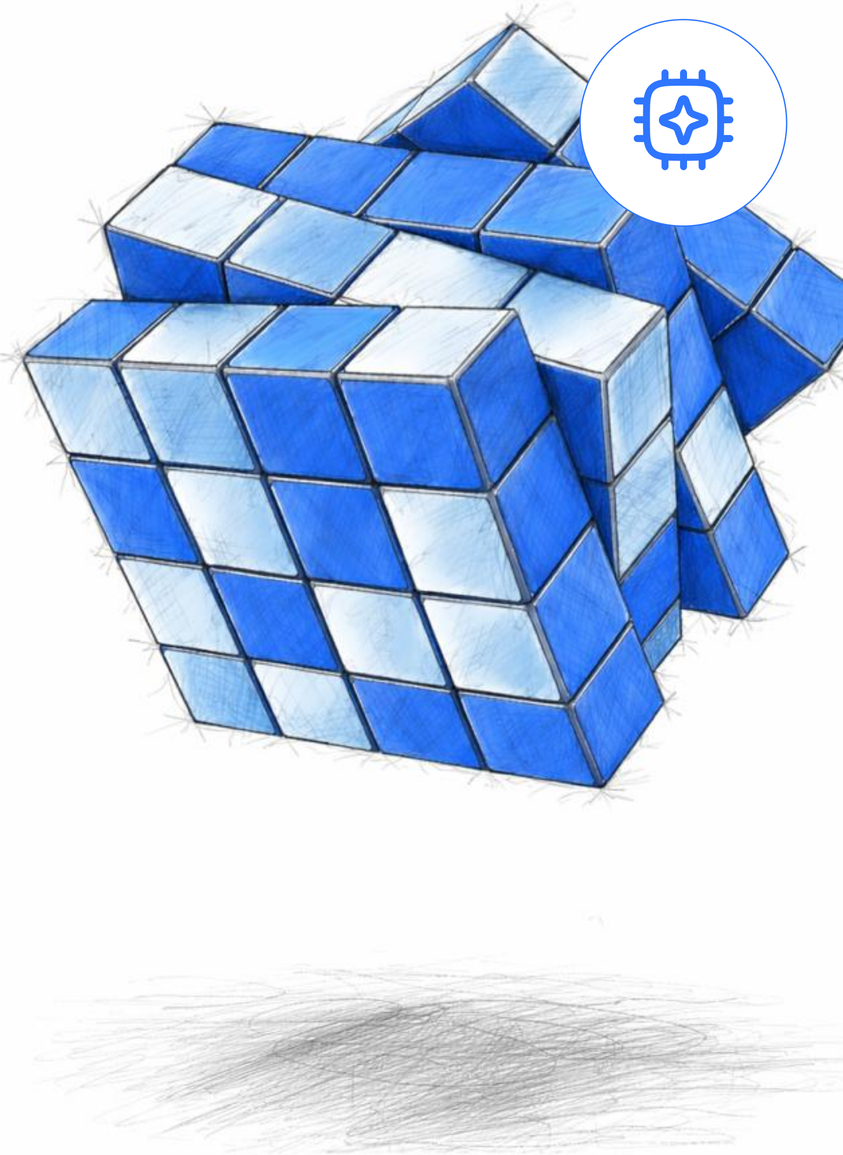
Wähle einen Meta Agent Run

Wähle einen Run aus der Liste, um seine Timeline und weitere Details anzuzeigen.

Meta-Agenten: Die Roadmap

Rollout-Status

Schedule Line-Agent:	LIVE	**Auftragsbestätigungen**
Accounting-Agent:	LIVE	**Rechnungsstellung**
Nächste Releases:	Stammdaten-Agenten	Q3 2026
	Procurement-Agenten	Q4 2026
	Monatlich neue Meta-Agenten für alle Produkte	
Verfügbarkeit:	Schrittweise pro Kunde aktivierbar	



blp connect

Wofür setzen wir Agenten ein?

1. Standard
2. Custom AI
3. Meta-Agenten
4. Standard-Erweiterungen
5. Unbound

02

blp connect

Vielen Dank