HARNESSING THE POWER OF AI IN RESEARCH AND DATA SCIENCE: TOOLS, TECHNIQUES, AND LIVE DEMOS

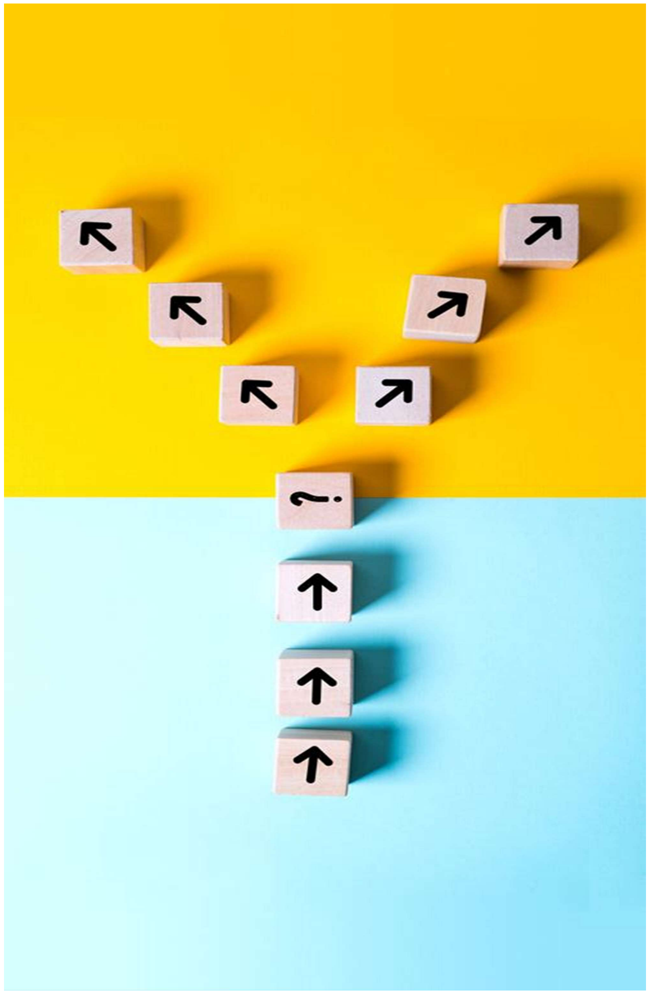
Minha Hwang

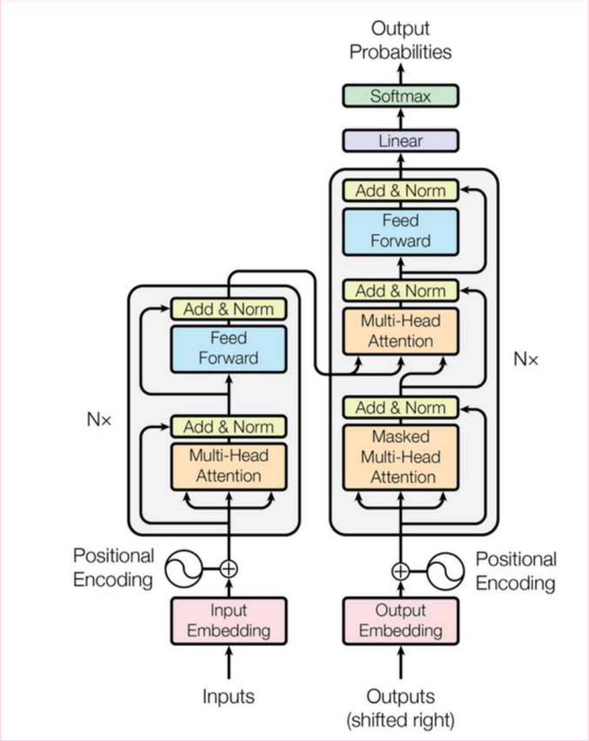
AGENDA

AI IN RESEARCH

AI IN DATA SCIENCE

AI TOOLS





AI IN RESEARCH

AI IN RESEARCH (1/2)

Research Process	**What Can Be Used**
Problem Formulation	Brainstorming with a strong reasoning model (e.g., GPT-o3)
Literature Review and Reference	**Deep Research**:Your RA
Research Design and Method • Paper to Code: Reproducing Past Research • Simulation Study Code Generation	AI Chatbot with strong coding and reasoning capabilities: GPT-o3, Claude Sonnet 4.0
Data Collection and Acquisition • Synthetic Data Generation from Prompt Engineering • Open-Source Data Collection	AI Chatbot with strong reasoning capabilities: GPT-o3, Claude Sonnet 4.0 **Deep Research**:Your RA

AI IN RESEARCH (2/2)

Research Process	What Can Be Used
Data Analysis and Modeling	AI-assisted coding and data science agent: e.g., coding, EDA, interactive app
Interpretation and Insight Generation	AI Chatbot with strong coding capabilities: GPT-o3, Claude Sonnet 4.0
Writing and Visualization	Custom GPT ChatGPT - Creative Writing Coach
Paper Review and Critique	AI Chatbot with strong reasoning capabilities: GPT-o3, Claude Sonnet / Opus 4.0
Research Paper Evaluation • Publication Likelihood • Journal Outlet	Custom GPT ChatGPT - Research Paper Evaluation Framework

DEEP RESEARCH

	Key Characteristics
ChatGPT (Plus)	• **Best Deep Research** on the Market • Once you make a request, ChatGPT asks clarifying questions. • Usually takes 1 – 20 minutes to generate synthesized report with references • Useful for initial exploratory research, literature review, references, paper critique / review
Gemini (Paid)	• **2nd Best Deep Research** on the Market • Gemini shows "step-by-step research plan." You can modify this as you want. • Usually takes 1 – 20 minutes • Useful for similar tasks above. Integration with Google services (YouTube, Google Search) is a plus.
Grok 3 (Paid)	• Okay Deep Research on the Market • You can use this to complement / validate deep research from above two services. • Usually takes 1 – 20 minutes • Useful for similar tasks above.
Perplexity	• Free Deep Research on the Market • You can use this to complement / validate deep research from above two services. • Usually takes 1 – 20 minutes • Useful for similar tasks above. Less comprehensive compared to first two.
M365 Copilot	• Only Enterprise Deep Research on Market: Secure Enterprise Data from Knowledge Graph • Add "Researcher" on M365 Copilot • Need M365 Enterprise subscription • Usually takes 1 – 20 minutes

- Claude Opus/Sonnet 4.0 introduced Deep Research recently
- Open Deep Research: Hugging Face

LIVE DEMO

AI IN DATA SCIENCE

GEN AI DATA SCIENTIST VS. TRADITIONAL DATA SCIENTIST

	Gen AI Data Scientist	(Traditional) Data Scientist
Coding	• Leverages **coding and app development agents** such as GitHub Copilot, Windsurf, Rue, Cursor.ai, Replit, Lovable, v0, • Designs code at a high-level; **Coding in English** with step-by-step plans • Focuses on efficiency by utilizing AI chatbot assistant and agentic development	• **Writes code end-to-end** in a familiar programming language. • Invests **significant time learning specific languages, libraries, and frameworks**. • Seeks solutions through **peer consultation** or resources like **Stack Overflow**.
Data analysis	• Utilizes **conversational data science agents** for exploratory data analysis (EDA). • **High-level design** of data pipeline and analysis plan • Uses **AI-assisted visualization tools** for efficient chart and graph creation. • Handles **unstructured data (text, image, voice)** with AI-powered tools.	• Either skips EDA or **relies on expert knowledge**. • Writes the entire data analysis pipeline **manually**. • Spends **significant time fine-tuning visualizations** (fonts, colors, layout). • Primarily focuses on **tabular (numerical) data**.
Focus	• Emphasizes **model evaluation and data curation**, including synthetic data generation. • Works with pre-trained models as a service, enabling limited adaptation. • Utilizes pre-trained models from platforms such as Hugging Face, Azure AI Foundry, Google AI Studio, and AWS Bedrock.	• Concentrates on **model training**, including relevance engineering and recommendation algorithms. • Develops models from scratch, focusing on model architecture, hyper-parameter optimization and training recipes.
Key Skills	• Strong **problem definition and problem-solving abilities**. • Structured **planning and efficient execution**. • Excellent **communication skills**. • Deep **understanding of AI mechanisms and functionalities**.	• Strong **technical expertise** in model training, statistics, and machine learning/deep learning (ML/DL). • **Limited emphasis on soft skills** and communication.

DATA SCIENCE AGENT

Microsoft 365 Copilot

- Add "Analyst" Agent (New)
- Upload and Analyze (Size Limit)
- Safe for Enterprise Data

Claude

BY ANTHROP\C

- Upload and Analyze (Size Limit)
- Only for Public / Fake Data

ChatGPT

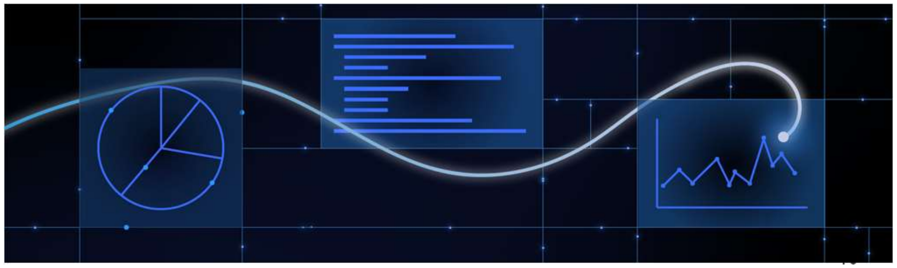
- Upload and Analyze (Size Limit)
- Only for Public / Fake Data

Gemini

- Integrated in Colab Notebook
- Only for Public / Fake Data

In addition, **use agentic code IDEs**

- GitHub Copilot (Agent mode)
- Cursor
- Windsurf



WHAT CAN YOU DO? (1/5)

Task	Prompt	Output
Data Cleaning	"Clean the dataset by removing rows with missing values and encoding categorical variables."	The agent returns a cleaned dataset with missing rows removed and categorical variables encoded using one-hot encoding.
Exploratory Data Analysis (EDA)	"Provide summary statistics, correlations across variables, histograms, and box plots for the 'sales' dataset."	The agent generates descriptive statistics (mean, median, standard deviation), correlations, and visualizations like histograms and box plots for the 'sales' dataset.
Feature Engineering	"Create a new feature representing the interaction between 'age' and 'income'."	The agent adds a new feature to the dataset by multiplying 'age' and 'income' columns, capturing their interaction effect.

Data analysis in **English** **: Prompting is All You Need**

- You just need to know "what to do"
- Do not need to look up PyPI or Stack Overflow for technical details and syntax

WHAT CAN YOU DO? (2/5)

Task	Prompt	Output
Model Building & Evaluation	"Train a logistic regression model to predict customer churn and evaluate its accuracy.	The agent trains a logistic regression model, outputs the accuracy score, and provides a confusion matrix to assess performance.
Natural Language Processing (NLP)	"Analyze the sentiment of customer reviews in the 'reviews' column."	The agent processes the text data, assigns sentiment scores (positive, negative, neutral) to each review, and summarizes the overall sentiment distribution.
Time Series Analysis	"Forecast the next 12 months of sales using the historical 'monthly_sales' data."	The agent fits a time series model (e.g., ARIMA), forecasts future sales, and plots the predicted values alongside historical data.

Data analysis in **English : Prompting is All You Need**

- You just need to know "what to do"
- Do not need to look up PyPI or Stack Overflow for technical details and syntax

WHAT CAN YOU DO? (3/5)

Task	Prompt	Output
Interactive Dashboard Creation	"Create an interactive dashboard showing key performance indicators (KPIs) such as conversion rate, click-through rate, and return on investment for the marketing campaign."	The agent develops a dashboard with visualizations like bar charts and line graphs, displaying KPIs such as conversion rate, click-through rate, and return on investment.
Data Summarization & Reporting	"Summarize the key findings from the quarterly sales data and generate a report highlighting trends and anomalies."	The agent analyzes the sales data, identifies significant trends (e.g., a 15% increase in Q2 sales), detects anomalies (e.g., a sudden drop in a specific region), and compiles a comprehensive report with visualizations and executive summaries.
Data Integration & Merging	"Merge the customer demographic dataset with the transaction history dataset to create a unified view for analysis."	The agent performs data cleaning, resolves discrepancies between datasets, and merges them based on common identifiers, resulting in a consolidated dataset ready for further analysis.

Data analysis in **English : Prompting is All You Need**

- You just need to know "what to do"
- Do not need to look up PyPI or Stack Overflow for technical details and syntax

WHAT CAN YOU DO? (4/5)

Task	Prompt	Output
Model Optimization	"Optimize the hyperparameters of the random forest model to improve prediction accuracy."	The agent conducts hyperparameter tuning using techniques like grid search or random search, evaluates model performance, and outputs the optimal parameters that enhance accuracy.
Text Classification	"Classify customer feedback into categories: Positive, Negative, or Neutral."	The agent processes the textual feedback, applies natural language processing techniques, and assigns each feedback entry to one of the specified categories, providing a summary of the distribution.
Anomaly Detection	"Identify any unusual patterns or outliers in the website traffic data over the past month."	The agent analyzes the traffic data, detects anomalies such as sudden spikes or drops in visits, and highlights potential causes or correlations with external events.

Data analysis in English **: Prompting is All You Need**

- You just need to know "what to do"
- Do not need to look up PyPI or Stack Overflow for technical details and syntax

WHAT CAN YOU DO? (5/5)

Task	Prompt	Output
Trend Analysis	"Analyze the trend of product returns over the last year and identify any seasonal patterns."	The agent examines the return data, identifies trends (e.g., increased returns during holiday seasons), and provides visualizations to illustrate these patterns.
Statistical Testing	"Conduct a t-test to determine if there's a significant difference in average purchase amounts between two customer groups."	The agent performs the t-test, calculates the p-value, and interprets the results to indicate whether the difference is statistically significant.
Data Pipeline Automation	"Automate the data ingestion and preprocessing steps for the daily sales data pipeline."	The agent sets up automated scripts or workflows that fetch daily sales data, clean and preprocess it, and store it in a designated database or data warehouse.

Data analysis in **English : Prompting is All You Need**

- You just need to know "what to do"
- Do not need to look up PyPI or Stack Overflow for technical details and syntax

LIVE DEMO

AI TOOLS

AI TOOLS (1/5)

Chatbots

- **CHATGPT**: HTTPS://CHATGPT.COM/ – **DEEP RESEARCH** (**RESEARCH ASSISTANCE**), **GPT STORE**, **OPERATOR** (AGENT), **AGENTS SDK**, **REASONING MODEL** (O1, O3, O1-PRO), IMAGE CREATION, **CODEX (CODING)**
- **MICROSOFT COPILOT**: HTTPS://COPILOT.MICROSOFT.COM/ – **OFFICE INTEGRATION** (E.G. **POWER-POINT SLIDE GENERATION** FROM WORD FILE, GITHUB COPILOT), FREE **REASONING MODEL** ACCESS (O3-MINI-HIGH), VISION, ACTION, GAMING, PODCASTS
- **CLAUDE.AI**: HTTPS://CLAUDE.AI/ – **FRONT-END WEB/APP DEVELOPMENT**, STRONG CODING, CREATIVE WRITING, **HYBRID-REASONING MODEL (SONNET 3.7), COMPUTER USE, TOOL INTEGRATION, DEEP RESEARCH** (**RESEARCH ASSISTANCE**), INTERACTIVE ARTIFACTS, **MCP TOOL INTEGRATION**
- **GOOGLE GEMINI**: HTTPS://GEMINI.GOOGLE.COM/ – (TRUE) **MULTI-MODAL MODEL**, GOOGLE SERVICE INTEGRATION E.G. YOUTUBE), **REASONING MODEL** WITH **TRANSPARENT INTERMEDIATE STEPS** (THINK EXPERIMENTAL), **DEEP RESEARCH**
- **DEEPSEEK-R1**: HTTPS://DEEPSEEK.COM/ – OPEN-WEIGHT **REASONING MODEL** WITH INTERMEDIATE STEPS (CENSORING), **LOW-COST TOKENS**, DISTILLED MODELS
- **GROK 3**: HTTPS://GROK.COM/– **DEEP RESEARCH**, **REASONING MODEL,** X PLATFORM INTEGRATION, PERSONA, WORKSPACE, IMAGE CREATION
- **PERPLEXITY**: HTTPS://WWW.PERPLEXITY.AI/ – **ANSWER ENGINE**, MODEL CHOICES, **STRONG RAG**, **DEEP RESEARCH**, CUSTOM DISTILLED MODEL (SONAR), PERPLEXITY LABS (CODE GENERATION, MINI APPS, ASSETS)
- **LIQUD.AI:** HTTPS://PLAYGROUND.LIQUID.AI/– **NON-TRANSFORMER-BASED ARCHITECTURE – LIQUID NN**: FAST INFERENCE AND **ON-DEVICE FOCUS**, **MULTIMODAL**

AI TOOLS (2/5)

Coding

- CURSOR.AI: CURSOR - THE AI CODE EDITOR
- GITHUB COPILOT: GITHUB COPILOT · YOUR AI PAIR PROGRAMMER
- WINDSURF (AGENTIC IDE):WINDSURF EDITOR BY CODEIUM
- ROOCODE: ROO CODE – YOUR AI-POWERED DEV TEAM IN VS CODE
- CLINE (AGENTIC IDE): CLINE - AI AUTONOMOUS CODING AGENT FOR VS CODE
- AUGMENT CODE: HTTPS://WWW.AUGMENTCODE.COM/
- OPENAI CODEX: OPENAI CODEX | OPENAI
- CLAUDE CODE: HTTPS://DOCS.ANTHROPIC.COM/EN/DOCS/AGENTS-AND-TOOLS/CLAUDE-CODE/OVERVIEW
- CODY: HTTPS://CODDY.TECH/ (DUOLINGO FOR CODING)

Web / App Dev

- REPLIT: HTTPS://REPLIT.COM/
- LOVABLE: HTTPS://LOVABLE.DEV/
- V0 BY VERCEL:V0 BY VERCEL

Research Assistance

- CHATGPT DEEP RESEARCH PAID SUBSRIPTION): HTTPS://CHATGPT.COM/
- GEMINI DEEP RESEARCH (PAID SUBSCRIPTION): HTTPS://GEMINI.GOOGLE.COM/APP
- GROK 3 BETA DEEP RESEARCH: HTTPS://GROK.COM/
- PERPLEXITY DEEP RESEARCH: PERPLEXITY
- CLAUDE DEEP RESEARCH: HTTPS://CLAUDE.AI/
- HUGGING FACE OPEN-SOURCE DEEP RESEARCH: HTTPS://HUGGINGFACE.CO/BLOG/OPEN-DEEP-RESEARCH

Reference: Roo Code vs Cline:Which is Better for Coding?

AI TOOLS (3/5)

Image Gen

- CHATGPT GPT-4O IMAGE GENARATOR
- IFREE AI IMAGE GENERATOR - IMAGE CREATOR IN BING
- GEMINI AI IMAGE GENERATOR - FREE TEXT-TO-IMAGE CREATION TOOL
- MIDJOURNEY
- INSTRUCTPIX2PIX - A HUGGING FACE SPACE BY TIMBROOKS

Video Gen

- META MOVIE GEN
- VEO - GOOGLE DEEPMIND
- SORA
- VIDEO EDITOR: CAPCUT | ALL-IN-ONE VIDEO EDITOR & GRAPHIC DESIGN TOOL DRIVEN BY AI
- SLIDE TO VIDEO – AI STUDIOS: SLIDESHOW MAKER WITH AI | TURN TEXT TO VIDEO IN MINUTES

Music Gen

- MUSICGEN - ADVANCED AI MUSIC GENERATION
- MUSICLM - AI MODEL FOR MUSIC GENERATION

Speech

- WHISPER - SPEECH-TO-TEXT (STT): OPENAI/WHISPER: ROBUST SPEECH RECOGNITION VIA LARGE-SCALE WEAK SUPERVISION
- ELEVEN LABS - SPEECH-TO-TEXT (STT) , TEXT-TO-SPEECH (TTS), VOICE CLONING. VOICE CHANGE – FREE TEXT TO SPEECH & AI VOICE GENERATOR | ELEVENLABS

AI TOOLS (4/5)

Agent

- OPEN AI OPERATOR: INTRODUCING OPERATOR | OPENAI
- OPENAI AGENTS SDK: OPENAI AGENTS SDK
- LANG CHAIN AND LANG GRAPH - ORCHESTRATOR: LANGGRAPH
- MICROSOFT AUTOGEN - ORCHESTRATOR : HTTPS://GITHUB.COM/MICROSOFT/AUTOGEN
- CREWAI: HTTPS://GITHUB.COM/CREWAIINC
- SEMANTIC KERNEL - ORCHESTRATOR: MICROSOFT/SEMANTIC-KERNEL
- MAGNETIC-ONE: MULTI-AGENT SYSTEM - AUTOGEN: MAGENTIC-ONE — AUTOGEN
- MANUS – GENERAL AGENT: MANUS

Robotics Foundation Model

- LE ROBOT – HUGGING FACE: HUGGINGFACE/LEROBOT
- ISAAC-GROOT – NVIDIA: NVIDIA/ISAAC-GR00T N1.5
- GEMINI ROBOTICS: GEMINI ROBOTICS - GOOGLE DEEPMIND

Applications

- NOTEBOOK LM: Google NotebookLM | Note Taking & Research Assistant Powered by AI
- CAREER DREAMER: Explore Your Possibilities with Career Dreamer - Grow with Google
- HELIX: A Vision-Language-Action Model for Generalist Humanoid Control
- TYPEFACE: Typeface - Enterprise Generative AI Platform for Marketing & Content Creation
- SYNTHESIA – Text-to-video with Avatars: https://www.synthesia.io/

AI TOOLS (5/5)

Online Learning

- OPEN-SOURCE PRE-TRAINED MODELS: MODELS - HUGGING FACE
- LLM/GEN AI ONLINE SHORT COURSES: COURSES - DEEPLEARNING.AI
- MICROSOFT GEN AI FOR BEGINNERS: MICROSOFT/GENERATIVE-AI-FOR-BEGINNERS: 21 LESSONS
- HUGGING FACE LEARN: HUGGING FACE – LEARN
- OPENAI ACADEMY: OPENAI ACADEMY
- 5-DAY GEN AI INTENSIVE COURSE – KAGGLE/GOOGLE: 5-DAY GEN AI INTENSIVE COURSE WITH GOOGLE LEARN GUIDE | KAGGLE

APPENDIX