LLMs & Biomedical Sci.

Rational

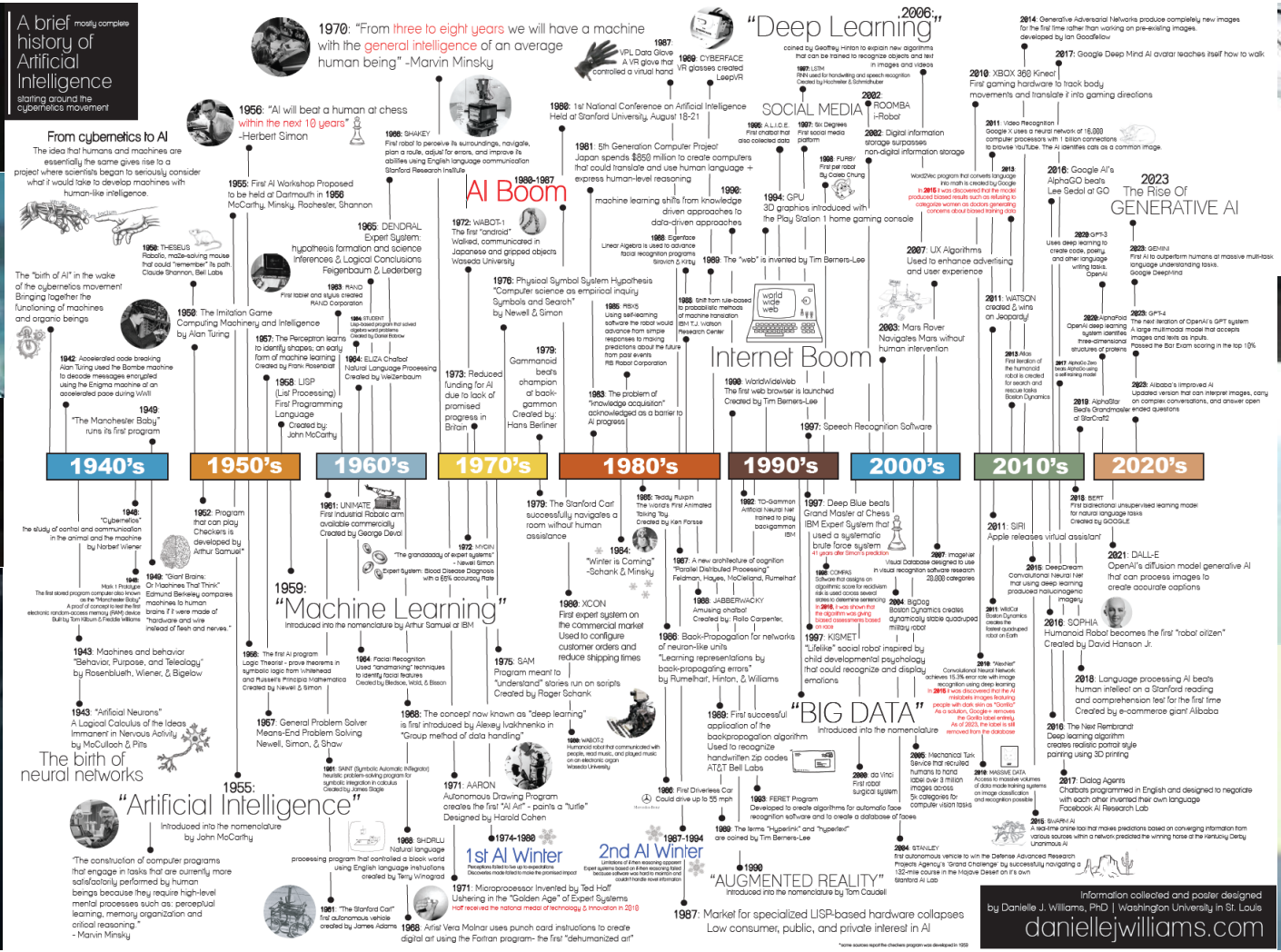
"Expand Big Data, Shared Equipment, and Core Service Capabilities"

-IBT Strategic Roadmap

Soft intro to AI

TEXAS A&M UNIVERSITY

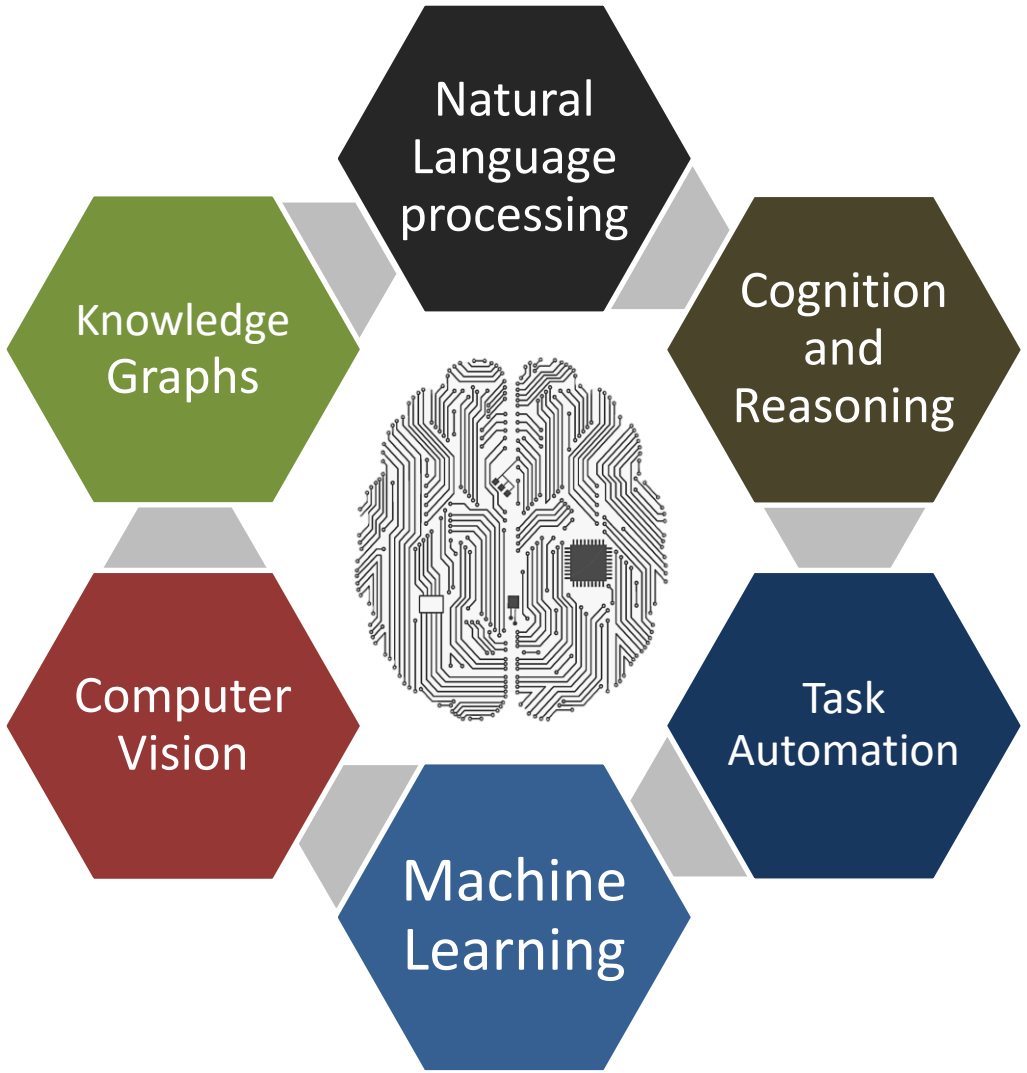
A brief mostly complete history of Artificial Intelligence starting around the cybernetics movement

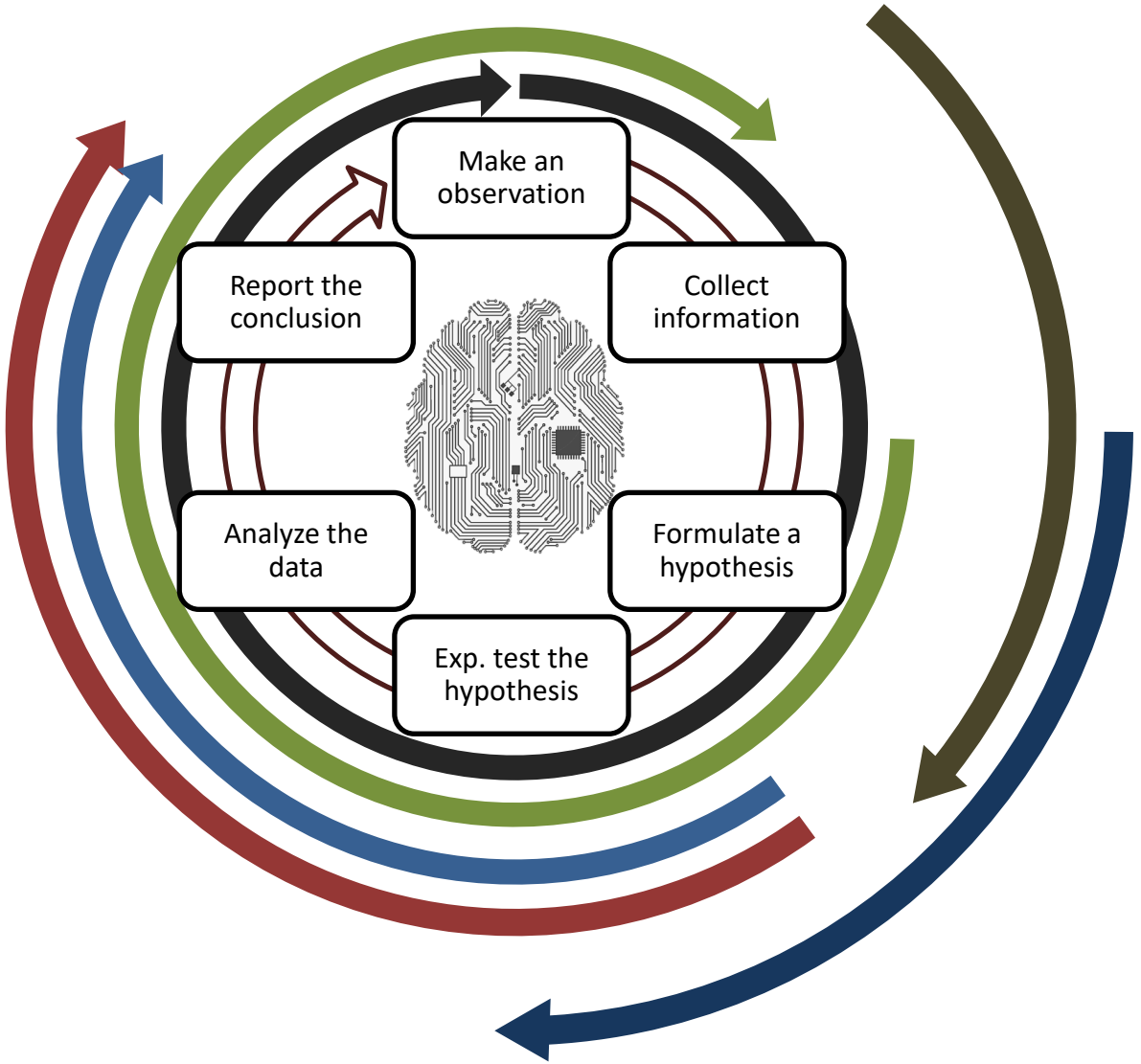


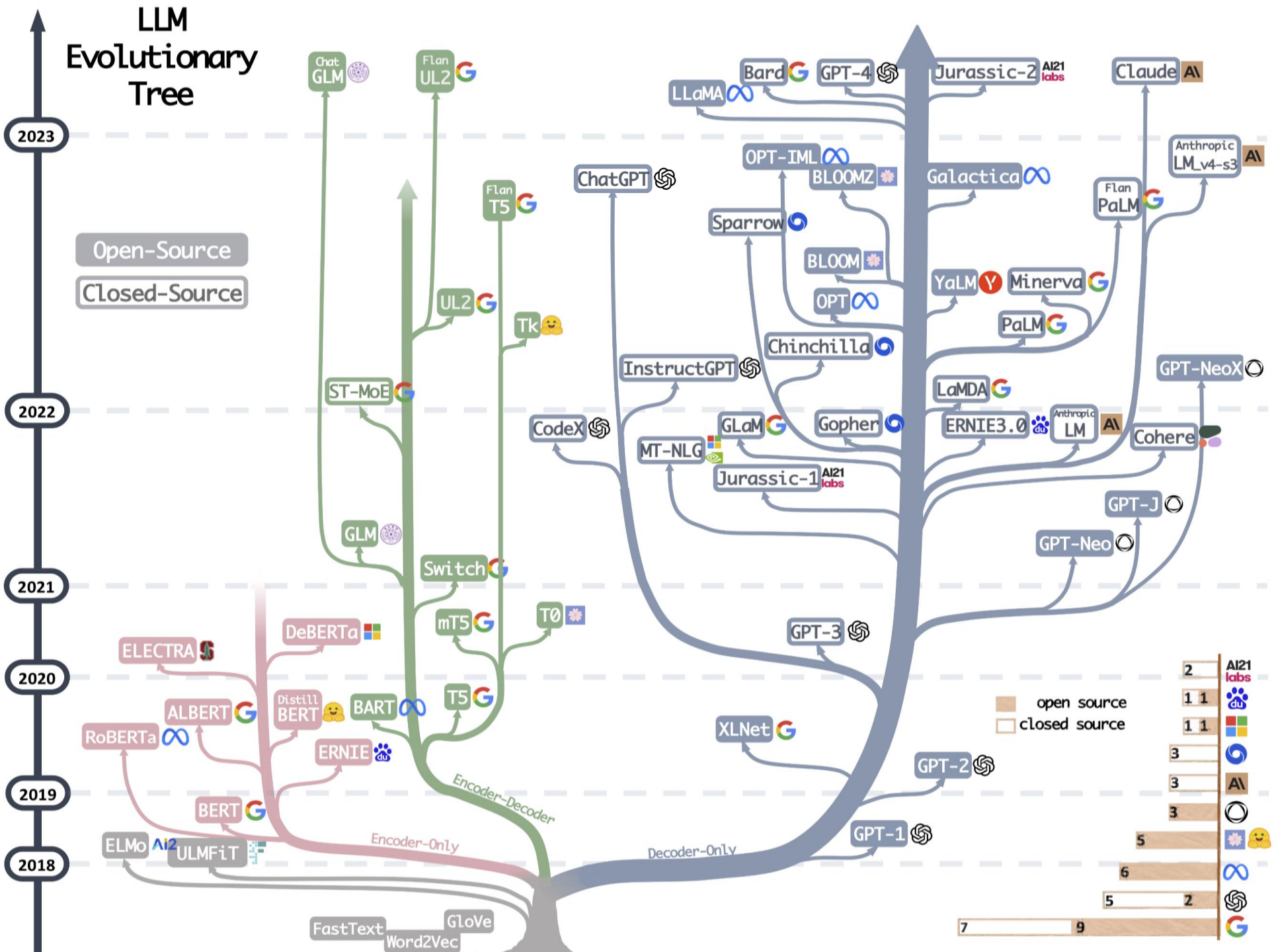
Information collected and poster designed by Danielle J. Williams, PhD | Washington University in St. Louis
daniellejwilliams.com

https://www.fastcompany.com/91129990/a-union-steps-into-the-spotlight

AI in the scientific method






LLM Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
Bard
GPT-4
Jurassic-2
Claude
LLaMA
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
Flan T5
ChatGPT
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
OPT
PaLM
UL2
Tk
Chinchilla
InstructGPT
GPT-NeoX
ST-MoE
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GLM
GPT-Neo
Switch
mT5
T0
DeBERTa
GPT-3
ELECTRA
ALBERT
Distill BERT
BART
T5
RoBERTa
XLNet
ERNIE
GPT-2
BERT
Encoder-Decoder
ELMo
ULMFiT
Encoder-Only
Decoder-Only
GPT-1
FastText
GloVe
Word2Vec
open source
closed source
AI21 labs
2
1 1
1 1
3
3
3
5
6
5 2
7 9

Foundational models

TEXAS A&M UNIVERSITY

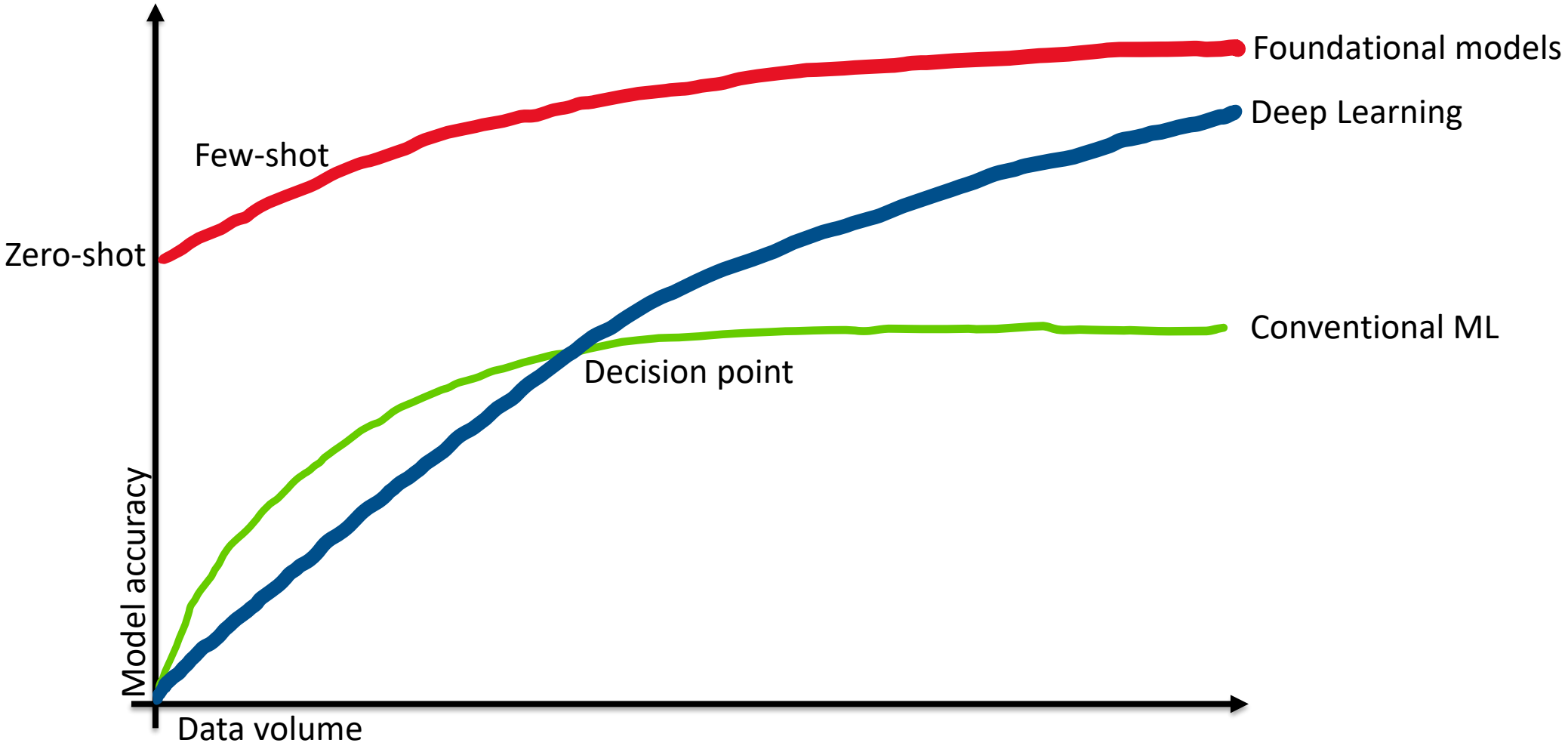
Broad Training Data: trained on extensive datasets, which require substantial computational resources. This training allows them to learn a wide range of tasks and skills during the initial phase.

Self-Supervision: Generally, use self-supervision techniques during training where labels or targets are generated from the data itself, rather than relying solely on human-labeled data.

Large Parameter Count: Typically contains at least billions of parameters to enable them to capture complex patterns and relationships in the data.

Applicability Across Contexts: Applicable across a wide range of contexts, can be secondarily fine-tuned for specific tasks with minimal adjustments, making them highly versatile.

Scaling models



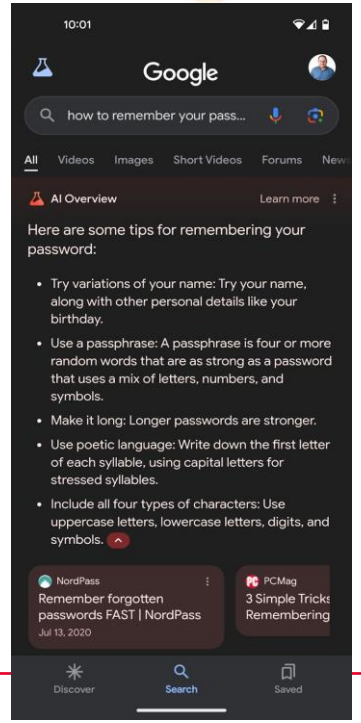
The dirty side of LLMs

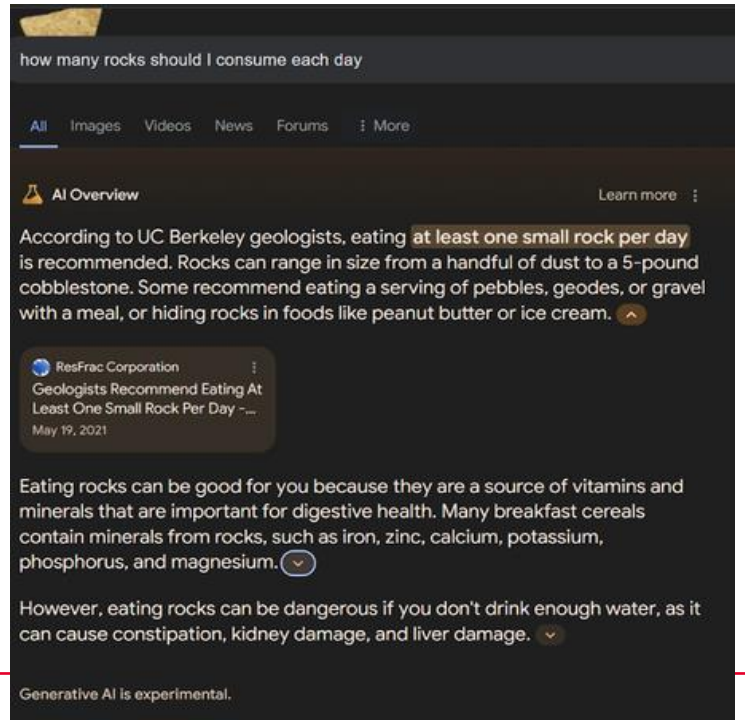
TEXAS A&M UNIVERSITY

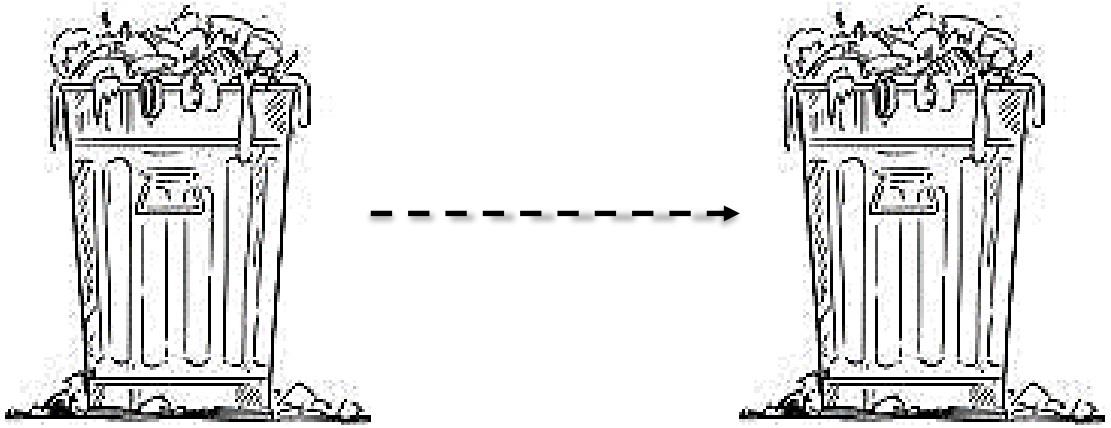
Thanks, but no thanks...

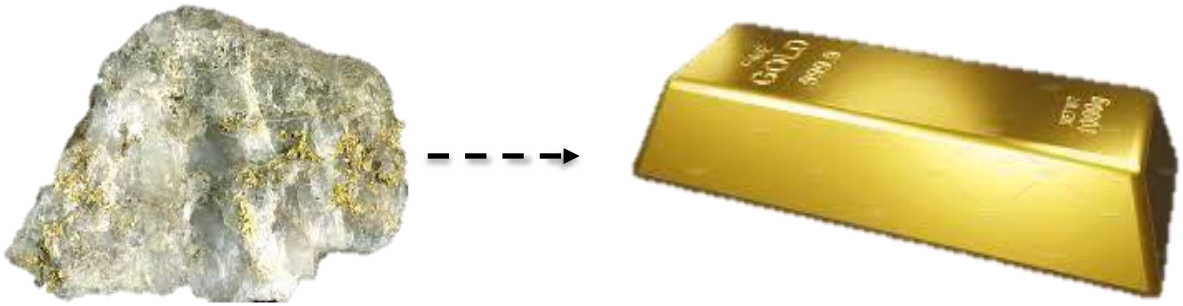
Here are some tips to help cheese stick to pizza:

- Add glue to the sauce: Mix about 1/8 cup of non-toxic glue into the sauce to make it tackier
- Use diced cheese: Diced cheese sticks more uniformly
- Let the pizza cool: After adding cheese, let the pizza cool for a few minutes so the cheese can settle and bond with the crust
- Freeze shredded cheese: Before adding cheese to the dough, freeze shredded cheese for 30 minutes









BioMed Foundational models

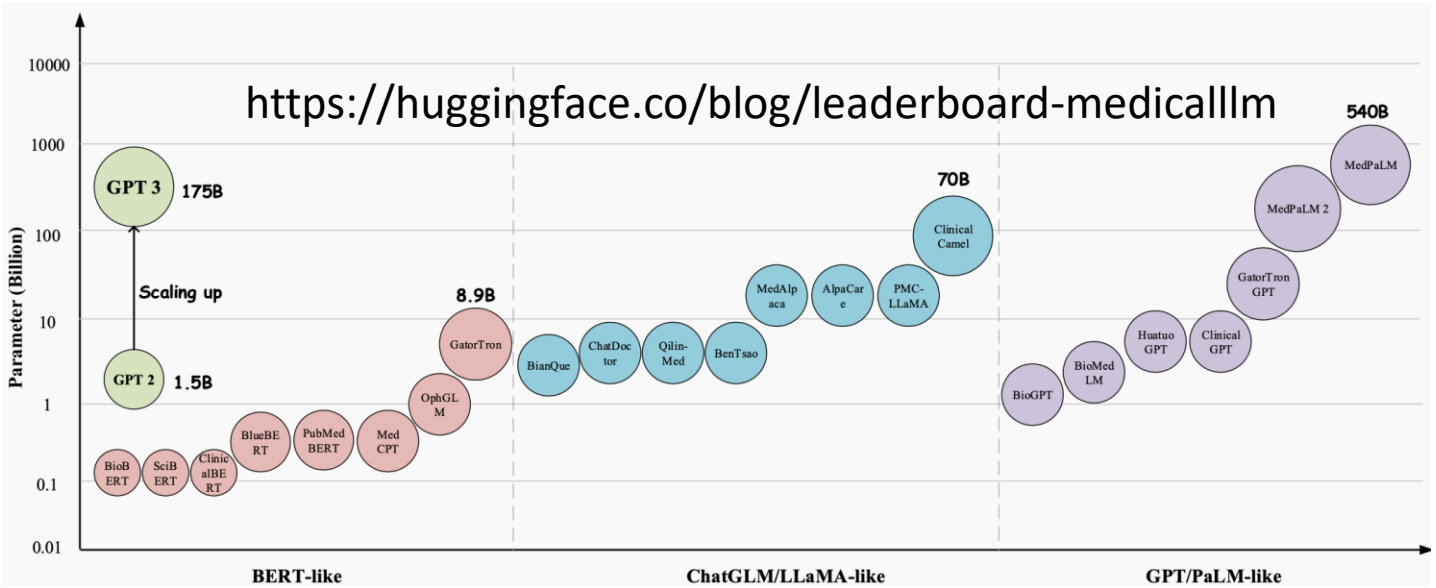
TEXAS A&M UNIVERSITY

Open Life Science AI

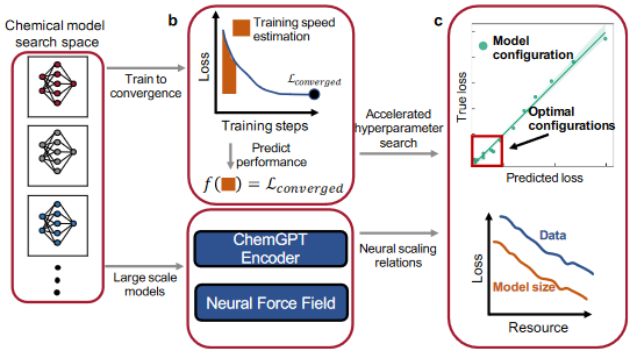
Edinburgh University of Edinburgh Natural Language Processing NLP

Open Medical-LLM Leaderboard

https://huggingface.co/blog/leaderboard-medicalllm

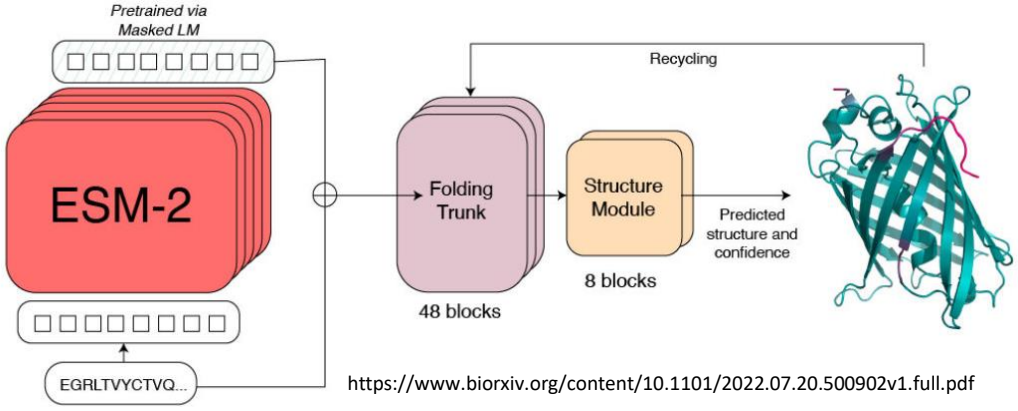


ChemGPT



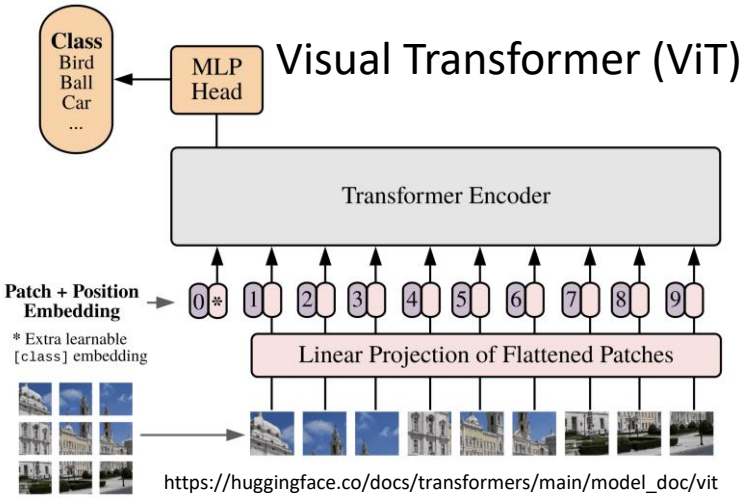
https://chemrxiv.org/engage/chemrxiv/article-details/627bddd544bdd532395fb4b5

ESM2: A large language model of protein structure



https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.full.pdf

Visual Transformer (ViT)



https://huggingface.co/docs/transformers/main/model_doc/vit

Prompt engineering

Role-task-format (RTF)

Before-after-bridge (BAB)

Task-action-goal (TAG)

Context-action-result-example (CARE)

Role-input-steps-example (RISE)

CLEAR

CONCISE

CORRECT

Using prompt engineering

TEXAS A&M UNIVERSITY

|<System>|{Sys}|<User>|{Usr}|<Agent>|

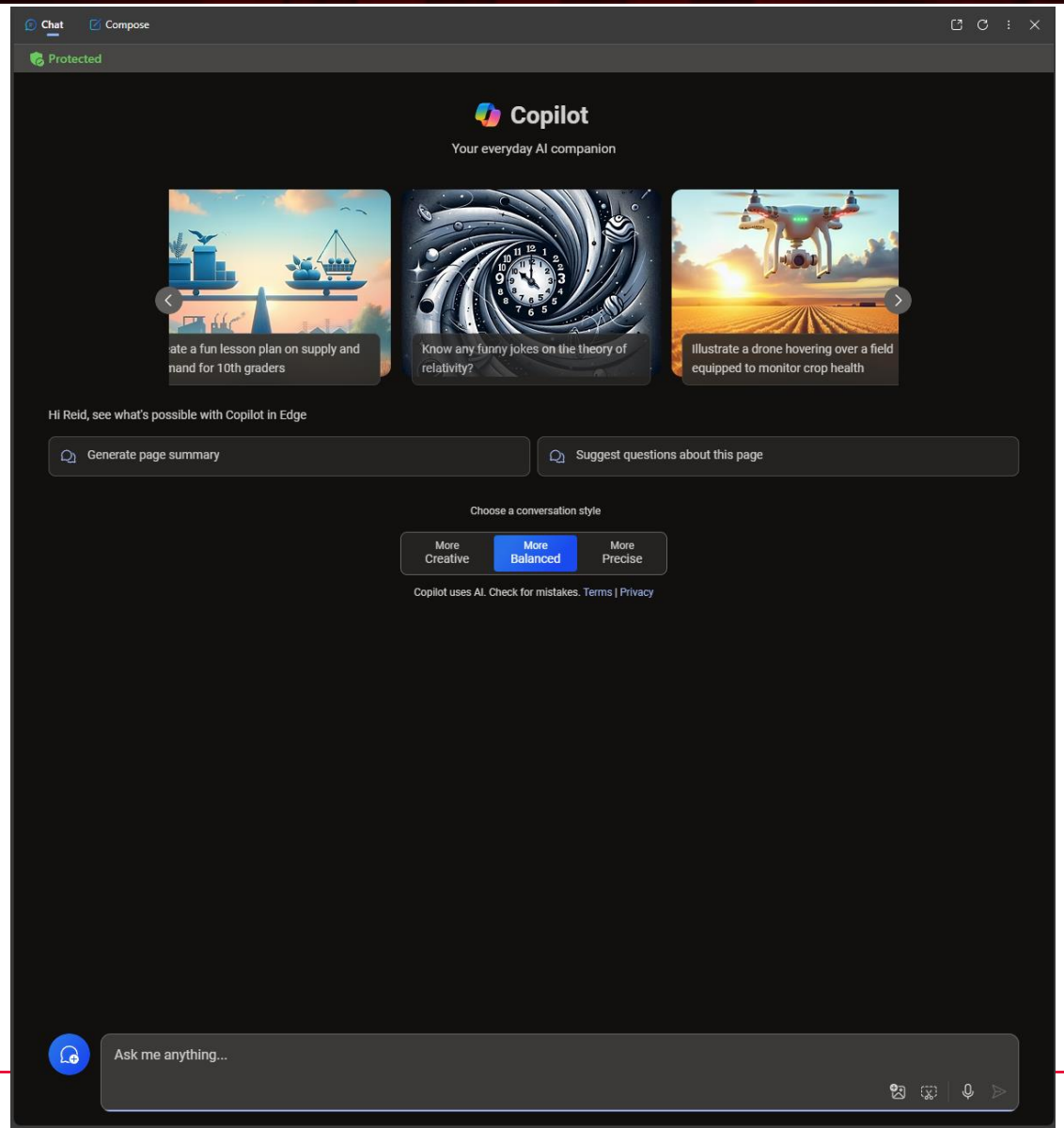
-the one ring

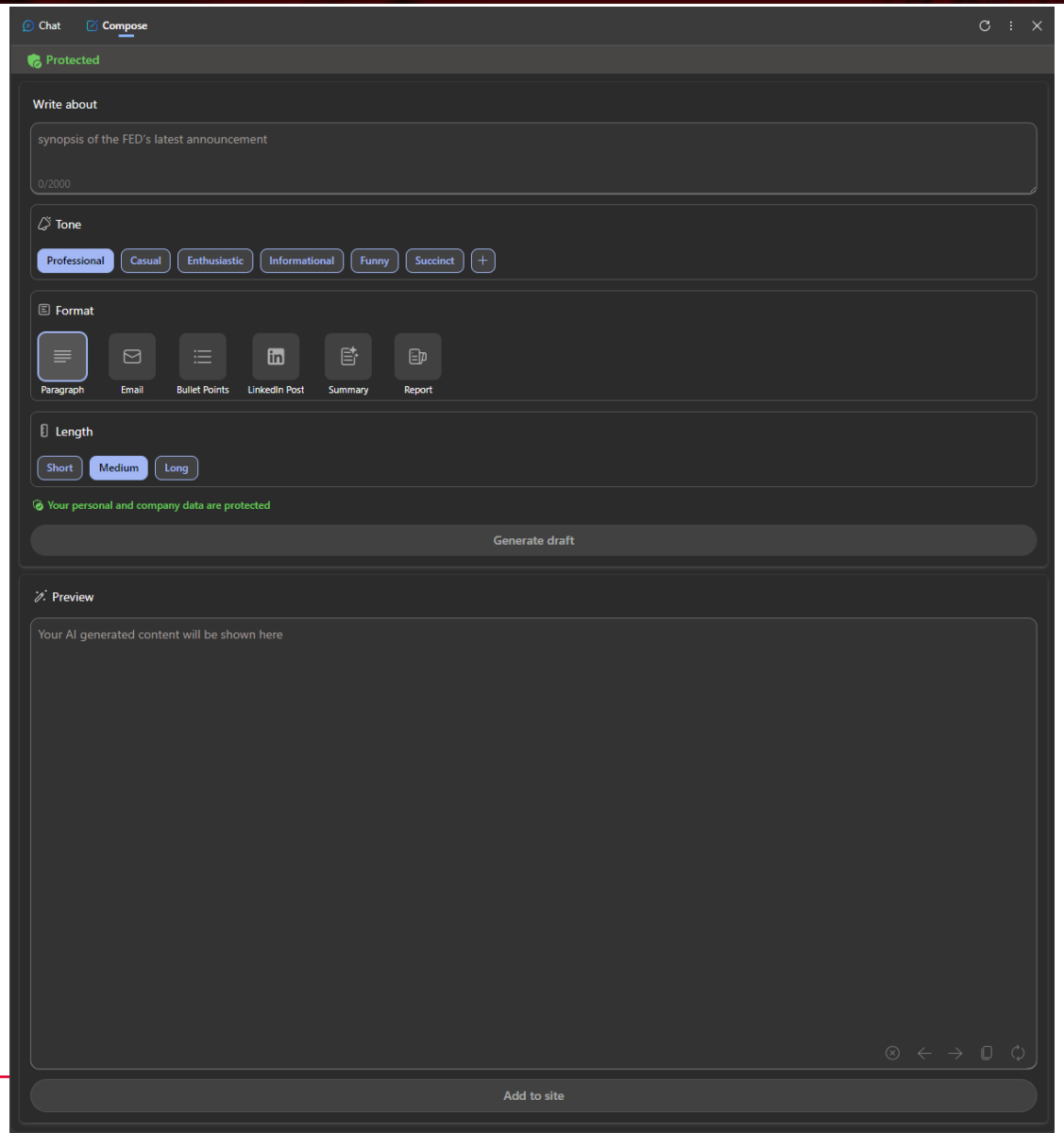
-Gollum

-The Mouth of Sauron

Microsoft Copilot

TEXAS A&M UNIVERSITY





Prompt Engineering (RTF)

TEXAS A&M UNIVERSITY

ROLE → TASK → FORMAT

Example 1: You are a professional scientific editor. Your task will be to review and edit the material that I will input in subsequent prompts. You will return the edited text in a block-text paragraph formats.

Example 2: You are a highly critical but fair reviewer. Your task will be to provide critical feedback on the text that I provide and identify any logical fallacies or internal inconsistencies. Your response should be in the form of itemized bullet points

Prompt Engineering (TAG)

TEXAS A&M UNIVERSITY

TASK → ACTION → GOAL

Task: To write anticipated results, limitations, pitfalls, and alternative approach section for a grant application

Action: Identify the rationale and hypothesis from the materials to be provided. Critically evaluate the experimental strategies, data collection methods, and overall approach using internet searches. Provide feedback on any logical facilities, alternative approaches, or pitfalls and how these risks may be mitigated.

Goal: To de-risk the experimental procedures being proposed in this grant application and ensure that any and all results will further the mission of the research effort.

Prompt Engineering (RISE)

TEXAS A&M UNIVERSITY

ROLE → INPUT → STEPS → EXAMPLE

System = "You are a knowledgeable biomedical researcher extracting key information from unstructured literature. The input, provided below is in paragraph format. The first task will be to perform named entity recognition of the most important content. Second, convert the most relevant terms into a JSON array with the following column headers: type, value, and relevance. For example, from the sentence, “Doxorubicin is a potent inhibitor of cancer growth” you generate the following json array {“type”:”drug”, ”value”:”doxorubicin”, ”relevance”:”growth inhibitor”}”

User = …Paragraph from literature about some drug related research…

Input = |<System>|{System}|<User>|{User}|<Agent>|

User prompt

TEXAS A&M UNIVERSITY

Synthesis and structure–activity relationship studies of MI-2 analogues as MALT1 inhibitors

Guolin Wu [1], Haixia Wang [2], Wenhui Zhou [3], Bihua Zeng [1], Wenhui Mo [1], Kejie Zhu [1], Rong Liu [4], Jia Zhou [5], Ceshi Chen [6], Haijun Chen [7]

Affiliations + expand

PMID: 29751989 DOI: 10.1016/j.bmc.2018.04.059

Abstract

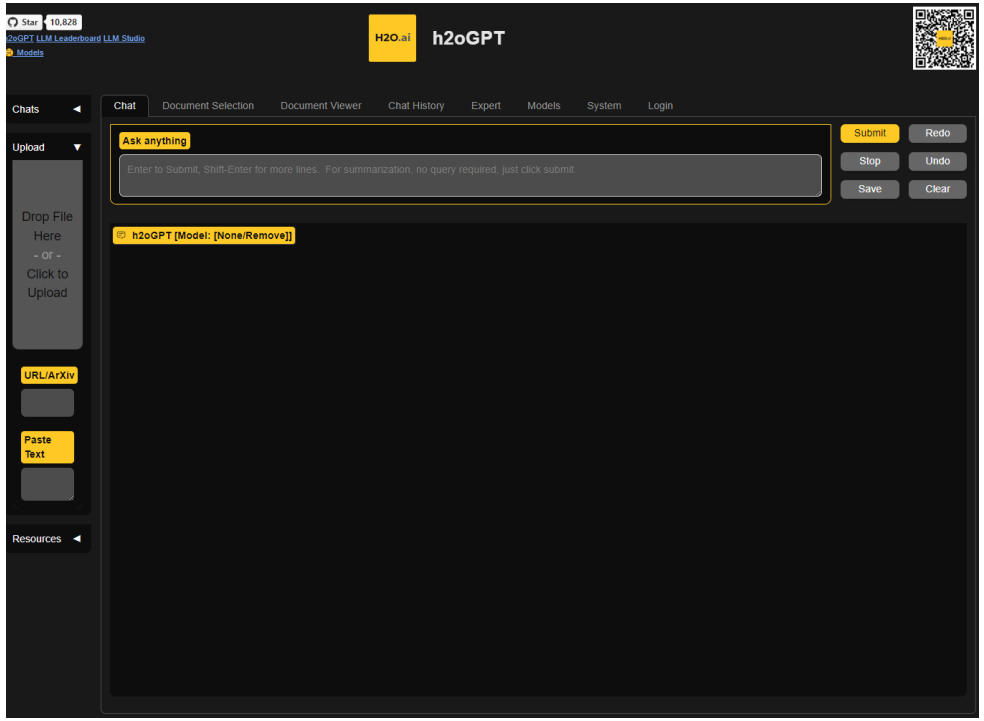
Recent studies revealed that MALT1 is a promising therapeutic target for the treatment of ABC-DLBCL. Among several reported MALT1 inhibitors, MI-2 as an irreversible inhibitor represents a new class of ABC-DLBCL therapeutics. Due to its inherent potential cross-reactivity, further structure-activity relationship (SAR) study is imperative. In this work, five focused compound libraries based on the chemical structure of MI-2 are designed and synthesized. The systematic SARs revealed that the side chain of 2-methoxyethoxy has little impact on the activity and can be replaced by other functionalized groups, providing new MI-2 analogues with retained or enhanced potency. Compounds 81-83 with terminal hydroxyl group as side chain displayed enhanced activities against MALT1. Replacement of triazole core with pyrazole is also tolerant, while structural modifications on other sites are detrimental. These findings will facilitate further development of small-molecule MALT1 inhibitors.

[
{"type": "target", "value": "MALT1", "relevance": "therapeutic target for ABC-DLBCL"},
{"type": "disease", "value": "ABC-DLBCL", "relevance": "disease targeted by MALT1 inhibitors"},
{"type": "drug", "value": "MI-2", "relevance": "irreversible inhibitor of MALT1"},
{"type": "chemical_group", "value": "2-methoxyethoxy", "relevance": "side chain with little impact on activity"},
{"type": "compound", "value": "Compounds 81-83", "relevance": "enhanced activities against MALT1 with terminal hydroxyl group as side chain"},
{"type": "chemical_core", "value": "Triazole", "relevance": "original core structure"},
{"type": "chemical_core", "value": "Pyrazole", "relevance": "tolerant core structure replacement"},
{"type": "chemical_group", "value": "Hydroxyl group", "relevance": "enhances activity against MALT1"}
]

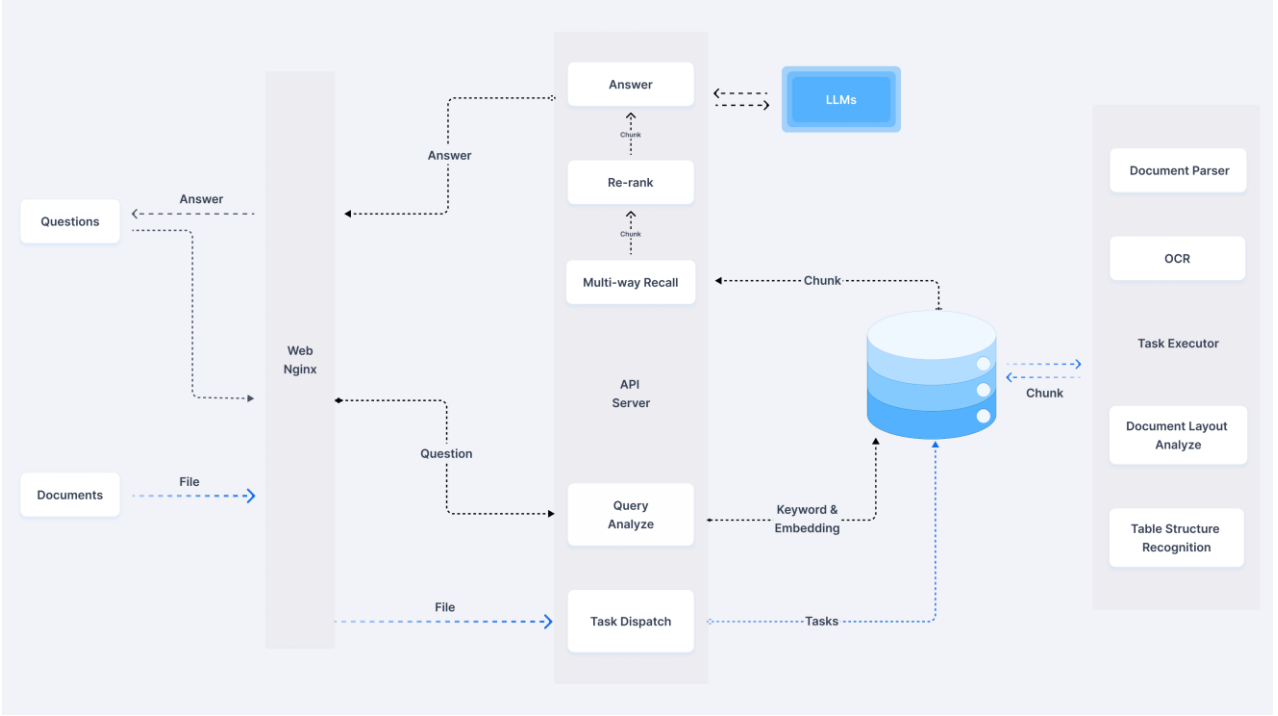
Retrieval Augmented Generation

TEXAS A&M UNIVERSITY

[h2ogpt: Private chat with local GPT with document, images, video, etc.]



[RAGFlow is an open-source RAG (Retrieval-Augmented Generation) engine based on deep document understanding.]



DGE analysis with ChatGPT+PE (Zero-shot)

I will give you a list of human gene transcripts. You will search the internet for the gene list and perfrom term enrichment on what you find, i.e. tell me what the commonalities are in their function. Make use of classification hierarchies when you do this. Include the top 25 terms that represent the biolgical processes associated with the complete gene list. Also include a hypothesis of the underlying biological mechanism or pathway.

Provide results in the format:

{{SUMMARY_KEYWORD}}: <high level summary>

{{MECHANISM_KEYWORD}}: <mechanism>

{{ENRICHED_TERMS_KEYWORD}}: <term1>; <term2>; <term3>

Here are the up regulated genes: TNFRSF12A, SLC41A2, HIST1H4E, ARHGAP40, APOBEC3B, HIST1H4B, ITM2B, TMBIM6, HIST1H1E, ARPC3, STXBP3, HIST1H2BG, SLC7A3, CALR, DNAAF2, WNK4, RNPS1, CCDC8, LIMA1, NCF4, NHLRC4, NEURL3, FCMR, KCTD5, BTN3A2, RBM23, BANF1, XRCC6, CKAP4, PSMA5, MRPS34, SIM1, LRTM1, HLA-DQA2, EPHB2, PDLIM1, HIST1H2BO, MSTN, ADAMTS18, RBX1, FDFT1, ITGAM, ILK, LRRN4, COMT, PMF1-BGLAP, IL1B, SCN5A, SPCS1, KRTAP29-1, ZFAND6, ICAM2, OMP, RDH16, C16orf91, PRB2, ARHGAP9, COLEC11, VAT1L, C11orf58, PDIA6, HIST1H4H, MAX, ADORA3, HIST1H4I, TRIM6, SCARB1, UBD, TRIM55, OAZ1, ,TUBA1B, SHISA5, TALDO1, OR1M1, UQCC2, GRM7, CTSE, SCGB3A1, CXCL10, TLR7, ABCC6, KRT9, PSMC3, OR51E2, ALG1, POLR2F, PSMD13, RTP4, HLA-F,MRPL58, EIF4G2, C19orf71, CALCA, NCF2, DYNLL1, MEIOB, OR5AU1, DES, MANF, SNCG, ALYREF, HSP90AB1, IL32, MUC7, HIST1H1C, ECI1, SERPINE3, RAN, ADGRE1, SOHLH2, SNX11, UBE2I, GSTM5, HNRNPCL3, PSME1, MTHFS, CCR3, FIBCD1, JSRP1, CTCFL, C22orf31, FBXO39, POLR3K, NOL7, PGA5 , HBQ1

Here are the down regulated genes: ZNF449, TNFAIP8L2-SCNM1, ZNF565, PCDHGA8, FAM114A2, OVOL3, KCNMB3, ZNF25, SLC4A3, PPFIA4, KMT2B, PCDHGA7, LOC100289561, MAPKBP1, ZNF709, TRMT10B, MCUB, VSTM5, APBB1IP, PCDHGA6, P2RY4, PTAR1, ELL3, ZNF335, LOC101927572, PBX4, OSER1, CDAN1, RHOBTB3, HRASLS, CCDC7, GAB2, CACNB1, ANKRD16, ZNF777, NKRF, ANGPTL7, PUS3, PNCK, LSM14B, SS18L1, GK5, PRELID2, TMEM52, ZNF3, INCA1, PLXNA3, ZNF461, DPH6, CNNM4, FBXO48, PABPC1L, PITRM1, FAM104B, LRFN3, SLC9B1, NFX1, NT5DC4, MAN2A1, PGBD4, USE1, ZNF248, UFSP2, GSDMC, GOLGA8A, ZNF287, ZNF789, DIDO1, NIPSNAP3B, CDK20, NEURL2, CCNL2, LPIN3, RRH, LNPK, CDC25B, RFESD, NR1D1, ZNF606, SLCO4A1, GRK5, CATSPER2, GARNL3, PLTP, SCAI, TAF4, DNAJC27, PTPDC1, OSBPL2, PCBD2, C20orf194, IGSF9B, ENO2, UCKL1, CDNF, TTC39B, ZNF416, NRBP2, RUBCN, GPC2, PIGO, ARHGAP33, DENND1A, SNAPC3, UVSSA, FAM58A, KCTD7, SLC20A1, ARID3A, HMCN1, ZNF586, ZNF438, C9orf129, MANEAL, REEP1, MVB12B, ZNF529, PLA2G4B, SAMD10, TNFRSF6B, LRWD1, RINT1, CCDC114, TIGD1, ZNF37A, RNF32, LPAR1, APLP1, SUCO, PDZD4, ENTPD1, L3MBTL1, PCDHGA2, ZNF813, PPP1R3D, LENG9, GPC4, ZNF250, PPFIA3, STBD1, RGS3, ZNF443, SLC9A8, CLIP2, MTHFD2L, ZNF566, ANO8, FAN1, MTMR10, EXOC8, TTC32, BRSK1, FAM167B, ZNF418, RELL2, ZNF417, DAND5, CATSPERG, SMURF1, SIN3B, RAB24, MRRF, C1orf74, REEP2, RNF145, BRAF, CCDC112, C20orf27, VAMP4, MAST3, NFS1, STAG2, PCDHGA1, PRRT3, C9orf3, ACTR3B, DNAJC2, ANKHD1, ENOX1, PLPPR2, FBXO41, MED26, ZNF550,KLF8, TTLL7, TIGD6, LENG8, KIF21B, ZNF397, INTS6L, FAM219A, GPRIN3, WDR31, ZNF112, NOM1, RABGAP1L, ADAL, MED7, ZNF786, ZNF805, ESR2, HERC3, AP4M1, ZNF497, ZBED8, UIMC1, MAP3K12, MAFK, ODF2, PRAF2, FEM1C, PER2, ZDHHC9, SLC16A5

Output

Summary for Upregulated Genes:

{{SUMMARY_KEYWORD}}: Heightened immune response and signaling pathways.

{{MECHANISM_KEYWORD}}: Activation of immune and signaling pathways in response to infection or inflammation.

{{ENRICHED_TERMS_KEYWORD}}: Immune response; Signal transduction; Transcription regulation; Cell adhesion; DNA repair; Protein ubiquitination; Cellular response to stress; Apoptotic process; Membrane transport; Cytoskeletal organization; Protein phosphorylation; Neuronal signaling; Cell cycle regulation; RNA processing; Chromatin organization; Protein catabolic process; G-protein coupled receptor signaling pathway; Calcium ion binding; Lipid metabolism; Cell differentiation; Wnt signaling pathway; MAPK cascade; Vesicle-mediated transport; Metabolic process; Zinc ion binding.

Summary for Downregulated Genes:

{{SUMMARY_KEYWORD}}: Reduced transcription and metabolic activity.

{{MECHANISM_KEYWORD}}: Downregulation of transcription and metabolic pathways possibly to conserve energy or resources during immune response.

{{ENRICHED_TERMS_KEYWORD}}: Transcription regulation; Signal transduction; Metabolic process; Cell adhesion; Zinc ion binding; Protein phosphorylation; Cellular response to stress; Apoptotic process; Membrane transport; Neuronal signaling; Immune response; Cell cycle regulation; DNA repair; Chromatin organization; RNA processing; Protein ubiquitination; Calcium ion binding; Lipid metabolism; G-protein coupled receptor signaling pathway; Wnt signaling pathway; MAPK cascade; Cell differentiation; Vesicle-mediated transport; Cytoskeletal organization; Protein catabolic process.

Follow-up prompt (Few shot)

This dataset is from cells responding to chemotherapy, given that information, how would you reframe your hypothesis and what drug classes would you propose for salvage therapy?

Reframed Hypothesis:

Cells exposed to chemotherapy exhibit upregulation of genes associated with stress response, immune response, and repair mechanisms, while downregulating genes involved in transcription, metabolic processes, and cell proliferation. This suggests an adaptive response to chemotherapeutic stress aimed at surviving the cytotoxic effects of the drugs. The upregulation of immune and stress response pathways indicates activation of survival mechanisms, while the downregulation of proliferative and metabolic pathways suggests a protective measure to minimize damage and conserve resources.

Proposed Drug Classes for Salvage Therapy:

Based on the observed gene expression changes, the following drug classes could be considered for salvage therapy to either enhance the efficacy of the chemotherapy or overcome resistance mechanisms:

1. **PARP Inhibitors**: Given the upregulation of DNA repair genes, inhibiting PARP (Poly ADP-Ribose Polymerase) could help in preventing repair of chemotherapy-induced DNA damage, thus enhancing cell death.
2. **Immune Checkpoint Inhibitors**: The upregulation of immune response genes suggests that the immune system is active. Immune checkpoint inhibitors could help in unleashing the immune system to attack tumor cells more effectively.
3. **Proteasome Inhibitors**: Upregulation of protein ubiquitination pathways suggests that proteasome activity is critical for the cells' response to chemotherapy. Proteasome inhibitors can disrupt protein degradation, leading to accumulation of damaged proteins and inducing cell death.
4. **HDAC Inhibitors**: Histone deacetylase (HDAC) inhibitors can modulate gene expression by altering chromatin structure, potentially reversing the downregulation of tumor suppressor genes and enhancing the efficacy of chemotherapy.
5. **Metabolic Modulators**: Drugs targeting altered metabolic pathways, such as glycolysis inhibitors or mTOR inhibitors, can exploit the downregulation of metabolic genes to further stress the cancer cells.
6. **Signal Transduction Inhibitors**: Given the changes in signal transduction pathways, inhibitors targeting specific signaling molecules like PI3K, AKT, or MAPK could disrupt survival signals in cancer cells.