



Intelligent Contextual Parsing and Synthesis of Disparate Information from Publications using Large Language Models

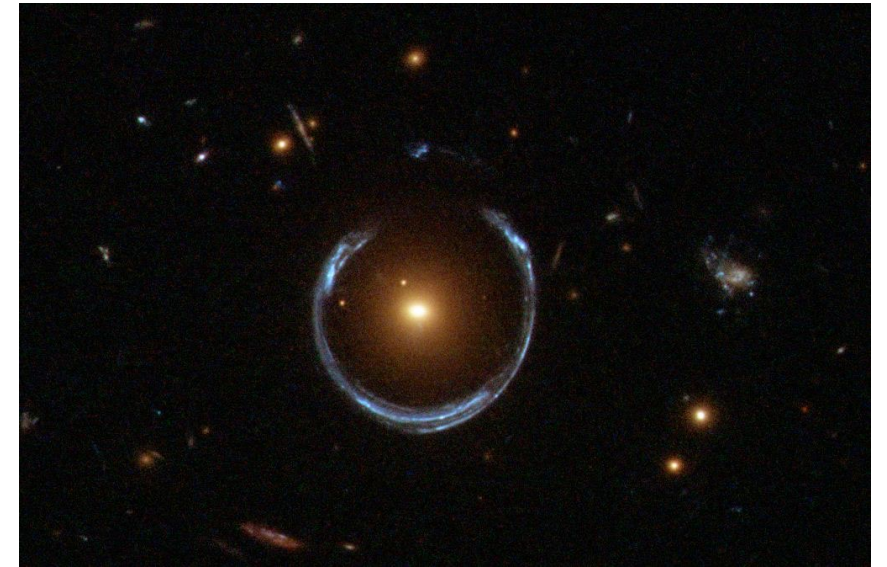
July 15, 2024

Jack Lightholder
jack.a.lightholder@jpl.nasa.gov

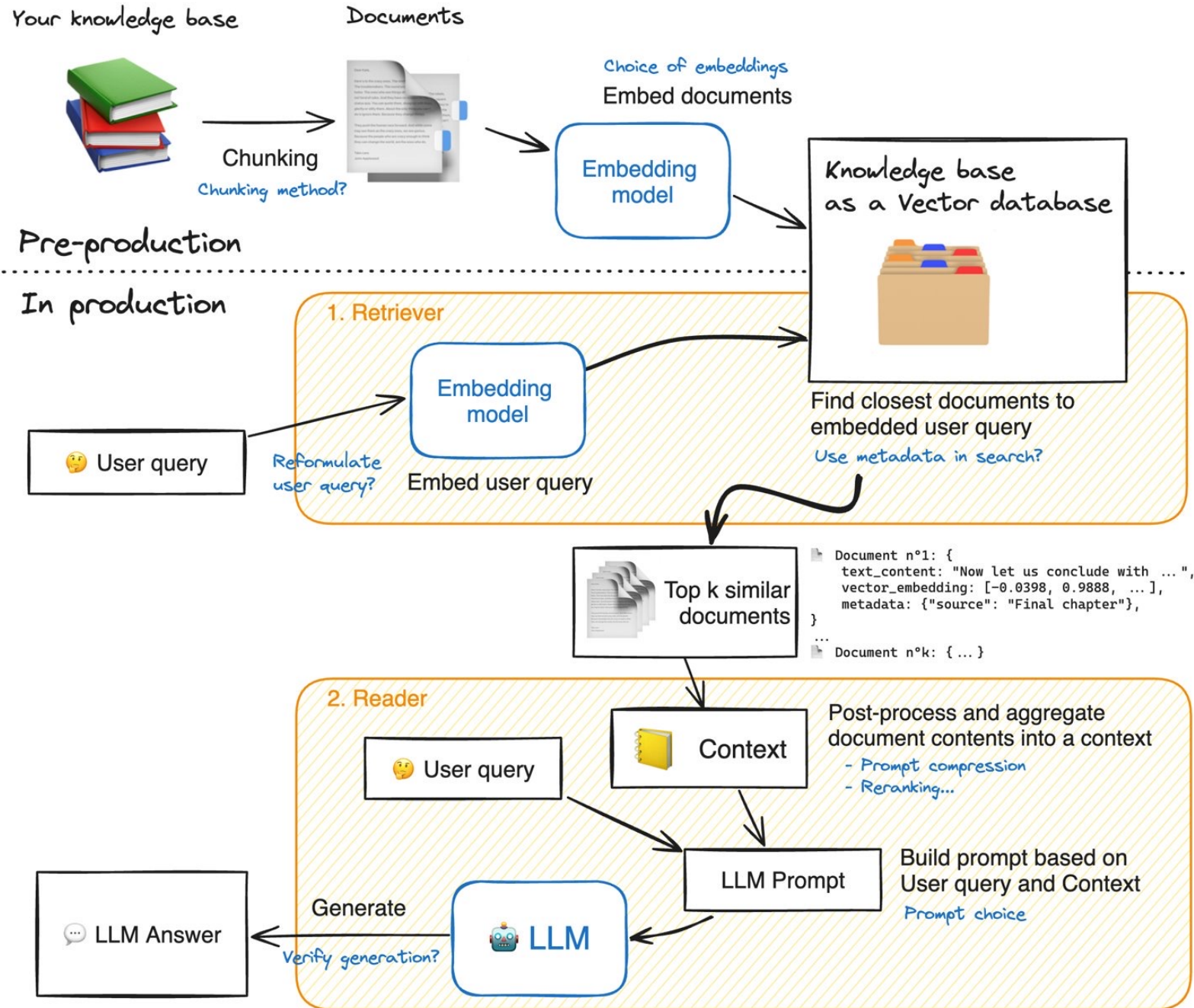
Leonidas Moustakas, Jake Lee, Mario Damiano

Opportunity

- Scientists spend thousands of hours curating databases of domain specific classes of data.
- New data is frequently buried in published literature, not easily machine readable or centralized.
- Gravitational lenses are going through a discovery renaissance thanks to missions like Euclid, JWST & Rubin.
- Lens database expected to go from 1,000s of known object to 100,000s in the next decade. **Manual curation of known lenses won't scale.**
- **We're developing a prototype LLM based system to facilitate parsing literature and extracting newly published gravitational lenses.**
- Leveraging data driven methods to identify classes of information commonly published about lens objects to inform storage architecture for future databases.
- Part of a larger effort to build a community hub for contributing newly found lenses



Architecture



Approach

- Curated and labeled gravitational lens literature
 - Identified papers representing common and challenging literature sources
- Deployed mixture of local and commercial models
 - Selection intended to span the options a scientist might consider utilizing
- Developed multiple prompt tree architectures to evaluate
- Developed JSON template to curate gravitational lens content populated by LLM
- Aggregated lessons learned for upcoming publication

System Name (1)	$m_{I_{814}}$ (Obs.) (2)	$\Delta m_{I_{814}}$ (extin.) (3)	R_{eff} (") (4)	q (5)
SDSS J0151+0049	19.816	0.049	0.665 ± 0.002	0.604 ± 0.001
SDSS J0747+5055	18.923	0.111	1.089 ± 0.002	0.737 ± 0.001
SDSS J0747+4448	19.417	0.066	0.924 ± 0.002	0.645 ± 0.001
SDSS J0757+4313	18.500	0.068	3.818 ± 0.009	0.572 ± 0.001
SDSS J0801+4727	19.911	0.103	0.499 ± 0.001	0.951 ± 0.002
SDSS J0821+3733	19.219	0.079	0.551 ± 0.001	0.743 ± 0.001
SDSS J0830+5116	19.332	0.087	0.969 ± 0.002	0.699 ± 0.001
SDSS J0837+4937	19.462	0.060	0.669 ± 0.001	0.473 ± 0.001
SDSS J0840+5051	19.984	0.043	0.357 ± 0.001	0.699 ± 0.001
SDSS J0841+5017	19.624	0.045	0.648 ± 0.001	0.939 ± 0.002
SDSS J0915-0055	18.518	0.070	1.219 ± 0.002	0.925 ± 0.001
SDSS J0941-0104	20.005	0.063	0.458 ± 0.001	0.568 ± 0.001
SDSS J0944-0147	19.965	0.067	0.478 ± 0.001	0.785 ± 0.002
SDSS J1016-0208	19.797	0.072	0.465 ± 0.001	0.816 ± 0.002
SDSS J1039-0014	19.345	0.094	0.812 ± 0.001	0.440 ± 0.001
SDSS J1117-0133	18.896	0.079	2.397 ± 0.006	0.938 ± 0.002
SDSS J1159-0007	19.463	0.049	0.958 ± 0.002	0.966 ± 0.002
SDSS J1215+0047	19.544	0.050	0.651 ± 0.001	0.684 ± 0.001
SDSS J1221-0220	18.942	0.055	0.710 ± 0.001	0.586 ± 0.001
SDSS J1221+3806	19.984	0.029	0.470 ± 0.001	0.838 ± 0.002
SDSS J1234-0241	19.269	0.064	1.054 ± 0.002	0.762 ± 0.002
SDSS J1318-0104	19.873	0.050	0.687 ± 0.002	0.761 ± 0.002
SDSS J1337+3620	18.603	0.023	2.034 ± 0.003	0.960 ± 0.001
SDSS J1344+3258	19.581	0.031	0.524 ± 0.001	0.746 ± 0.001
SDSS J1345-0129	21.877	0.071	1.000 ± 0.003	0.000 ± 0.001
SDSS J1349+3612	18.555	0.025	1.886 ± 0.003	0.743 ± 0.001
SDSS J1352+3216	19.514	0.024	0.579 ± 0.001	0.949 ± 0.001
SDSS J1452+3323	19.487	0.026	0.623 ± 0.001	0.836 ± 0.001
SDSS J1503+3225	20.118	0.032	0.769 ± 0.003	0.625 ± 0.002
SDSS J1522+2910	19.534	0.046	0.890 ± 0.002	0.579 ± 0.001
SDSS J1537+0220	19.682	0.111	0.386 ± 0.001	0.694 ± 0.001
SDSS J1541+1812	19.648	0.064	0.759 ± 0.002	0.755 ± 0.002
SDSS J1542+1629	18.580	0.054	0.726 ± 0.001	0.786 ± 0.001

Literature Curation

- Collected and evaluated literature known to contain gravitational lens data.
- Characterized commonality of reported parameters across publications to inform a future data management strategy.
- Identified short list of ‘easy’ and ‘hard’ papers to use as an evaluation data set.
- Isolate why ‘hard’ papers are difficult to successfully parse.
 - Object naming variations
 - Domain specific glyphs
 - Target data interwoven with irrelevant data
 - Complex table formatting
 - Poorly formatted source documents.

System Name (1)	Plate-MJD-Fiber (2)	z_L (3)
SDSS J015107.37+004909.0	3606-55182-0679	0.5171
SDSS J021214.80+002719.1	4236-55479-0603	0.5372
SDSS J074724.12+505537.5	3677-55205-0551	0.4384
SDSS J074734.75+444859.3	3676-55186-0581	0.4366

System Name (1)	m_{I814} (Obs.) (2)
SDSS J0151+0049	19.816
SDSS J0747+5055	18.923
SDSS J0747+4448	19.417
SDSS J0757+4313	18.500
SDSS J0801+4727	19.911
SDSS J0821+3733	19.219

Dropped precision in naming convention (same paper)

TABLE 5
BELLS GRADE-A STRONG LENS SIE MODEL PARAMETERS

System Name (1)	θ_E (") (2)	q_{SIE} (3)	$P.A.$ (°) (4)	N_S (5)	m_{814} (6)	μ (7)
SDSS J0151+0049	0.676	0.752	111.0	1	22.51	8.71
SDSS J0747+5055	0.754	0.641	4.9	2	21.46	2.95
SDSS J0747+4448	0.610	0.723	147.1	1	23.77	39.72
SDSS J0801+4727	0.491	0.891	41.1	2	22.07	3.82

Domain specific glyphs with references to in text descriptions

Name	RA	DEC	Selection	W1 - W2, W1	Gaia G	Date, Exp. Time	Outcome
J0003+4555	0.96401	45.92215	D+WISE	0.37, 17.95	17.61, 18.97	12 Sep, 600s	quasar+star
J0011-0845	2.83435	-8.76407	D+GMM/WISE	0.10, 17.89	20.31, 20.35	12 Sep, 1200s	lens, $z=1.70$
J0028+0631	7.09369	6.53195	S+GMM/WISE	0.16, 16.86	18.95	12 Sep, 1200s	lens, $z=1.06$
J0030-1525	7.56378	-15.41752	S+WISE	-0.09, 17.01	19.30	13 Sep, 1200s	quad lens, $z=3.36$
J0123-0455	20.84084	-4.93266	S+GMM/WISE	0.21, 17.28	20.29	12 Sep, 2100s	lens, $z=1.38$

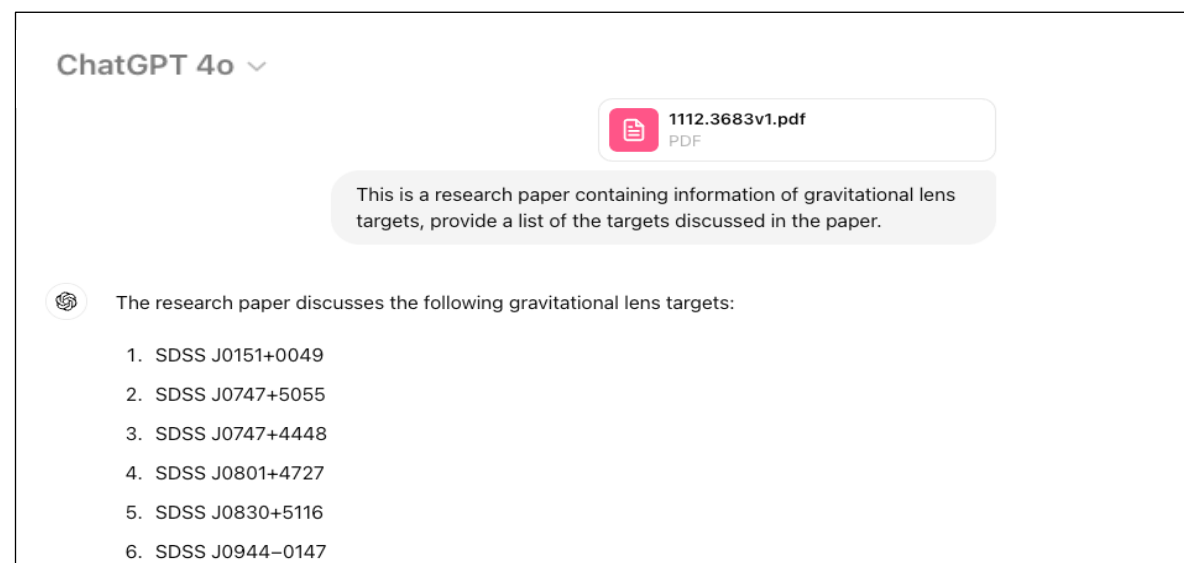
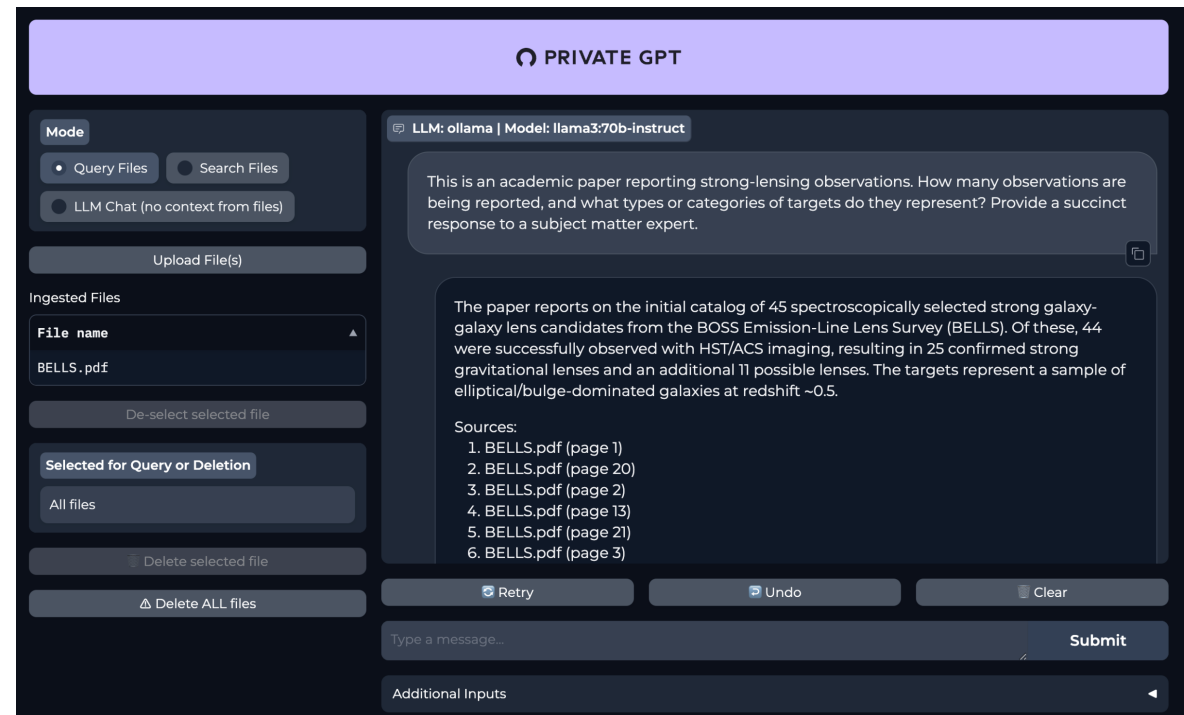
Data of interest interwoven with other data

name	b (")	PA_{SIE}	q_{SIE}	PA_{light}	q_{light}	χ^2_{gal}	χ^2_{images}	χ^2_{flux}	μ
J0011-0845	0.96 ^{0.97} _{0.95}	176 ¹⁷⁷ ₁₇₄	0.70 ^{0.73} _{0.68}	99 ¹³¹ ₅₅	0.86 ^{0.97} _{0.73}	0.07	0.19	0.03	5.0 ^{5.3} _{4.5}
J0028+0631	1.43 ^{1.44} _{1.42}	55 ⁵⁷ ₅₁	0.81 ^{0.83} _{0.79}	58 ⁶² ₅₄	0.86 ^{0.88} _{0.84}	0.07	0.19	0.02	4.2 ^{4.4} _{4.1}
J0030-1525†	1.08 ^{1.15} _{1.05}	170 ¹⁸² ₁₆₅	0.82 ^{0.95} _{0.33}	55 ⁵⁷ ₅₁	0.81 ^{0.83} _{0.79}	0.97	13.0	13.1	71 ⁸⁸ ₂₉

Complex table formatting

Model Selection & Deployment

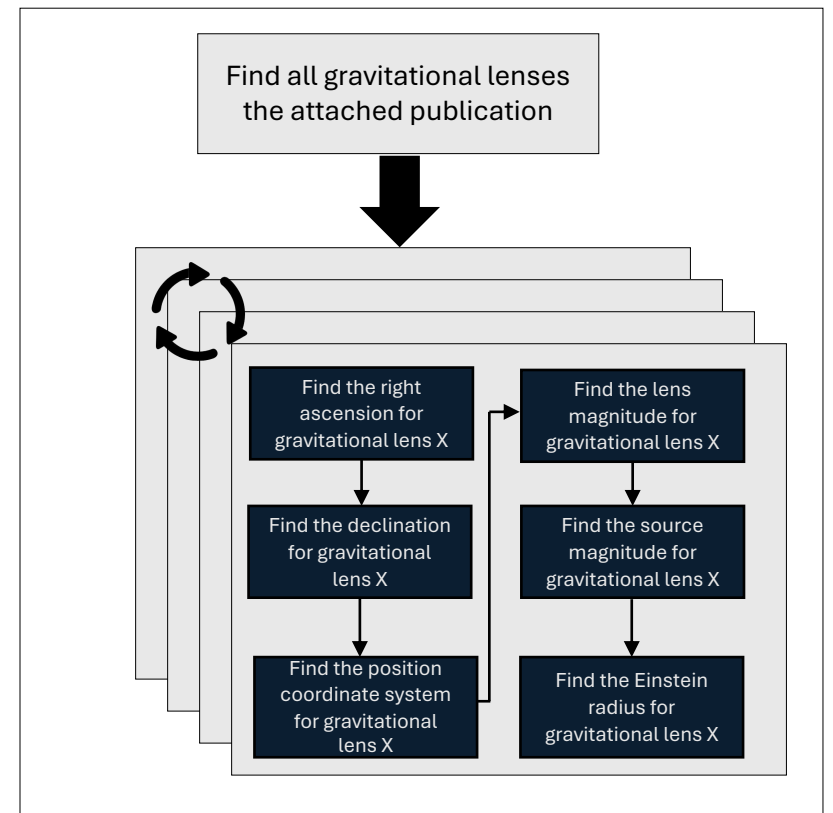
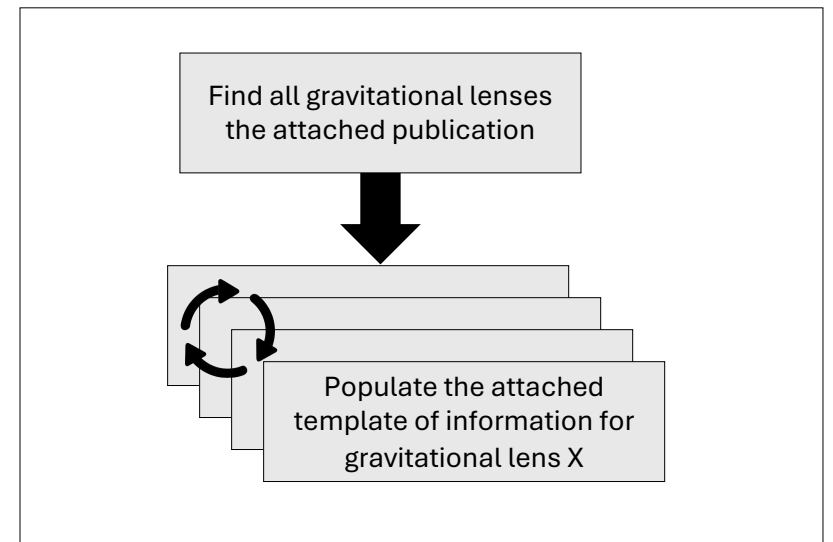
- Deployed local models using Private GPT.
- Deployed on high performance computing running A100 GPUs.
- Locally deployed models
 - Llama 2
 - Llama 3 (8B / 70B)
 - Mixtral 8x22b
- Purchased access to commercial models
- Ran using GUI terminals
 - Manual interaction
 - Variable context length
- Commercial API Models
 - GPT-4o (OpenAI)
 - Claude3 Opus (Anthropic)
 - Claude3 Opus (Perplexity)



Prompt Engineering

- Three step prompting approach
 1. Parse gravitational lenses contained in the provided paper
 2. Parse information about each lens contained in the paper
 3. Query published information missing from storage template
- Multiple prompting strategies
 - Request parsing of all template fields for a given gravitational lens in a single request (one shot)
 - Request parameters sequentially for a given gravitational lens (tree)









```
{  
  "system_name": "",  
  "discovery_date": "",  
  "alternate_name": "",  
  "kind_acronym": "",  
  "discovery_acronym": "",  
  "reference_identifier": "",  
  "ra_hrs": "",  
  "ra_mins": "",  
  "dec_coord": "",  
  "lensgrade": "",  
  "number_images": "",  
  "theta_e": "",  
  "theta_e_err": "",  
  "theta_e_quality": "",  
  "z_lens": "",  
  "z_lens_err": "",  
  "z_lens_quality": "",  
  "fluxes": "",  
  "mag_lens": "",  
  "mag_source": "",  
  "morphology": "",  
  "vett_status": "",  
  "released_status": "",  
  "hidden_status": "",  
  "lens_name": ""  
}
```



Curated Data Management

Follow model established by IPAC for exoplanets – Common values tracked by source

NASA EXOPLANET ARCHIVE
A SERVICE OF NASA EXOPLANET SCIENCE INSTITUTE

Source	Cadieux et al. 2024  	Kokori et al. 2023  	Lillo-Box et al. 2020  	Ment et al. 2019  
T_{eq} (K)	422±7	---	708.9 ^{+8.0} _{-7.8}	438±9
b	0.090 ^{+0.090} _{-0.060}	---	---	---
$S(S_{\oplus})$	5.3±0.4	---	---	6.16±0.37
$M_p(M_{\oplus})$	1.91±0.06	---	1.76 ^{+0.17} _{-0.16}	1.81±0.39
$M_p(M_{Jup})$	0.00601±0.00019	---	0.00554 ^{+0.00053} _{-0.00050}	0.00569±0.00123
e	<0.050	---	<0.274	<0.31
i (deg)	89.80 ^{+0.14} _{-0.19}	89.92 ^{+0.06} _{-0.09}	89.913 ^{+0.046} _{-0.049}	89.92 ^{+0.06} _{-0.09}
P (days)	3.777940±0.000002	3.7779329±0.0000028	3.77792±0.00003	3.777931±0.000003
ρ (g/cm ³)	5.1±0.4	---	6.07 ^{+0.81} _{-0.74}	4.7±1.1
$R_p(R_{\oplus})$	1.272±0.026	---	1.169 ^{+0.037} _{-0.038}	1.282±0.024
$R_p(R_{Jup})$	0.1135±0.0023	---	0.1043 ^{+0.0033} _{-0.0034}	0.1144±0.0021
K (m/s)	2.42±0.07	---	2.22±0.20	2.35±0.49
a (au)	0.0270±0.0005	---	0.02734±0.00054	0.02675±0.00070
δ (%)	0.290±0.009	---	0.252 ^{+0.016} _{-0.015}	---
T_{14} (hours)	1.13±0.02	---	1.100 ^{+0.025} _{-0.024}	---
T_c (days)	2458389.2939±0.0002	2458226.843969±0.000018	2458389.29383 ^{+0.00081} _{-0.00082}	2458226.843169±0.000026
a/R_*	---	26.57±0.05	27.53 ^{+0.62} _{-0.61}	26.57±0.05
R_p/R_*	---	0.05486±0.00013	---	0.05486±0.00013
$M_p \sin i (M_{\oplus})$	---	---	1.71±0.18	---
$M_p \sin i (M_{Jup})$	---	---	0.00538±0.00057	---

Tracked Parameters

Parameter Sources

System Level Considerations

- Must distill prompts to clear tasks which you want done at scale
 - Direct parsing of information easier than summarization/distillation of key points.
- Automatable APIs can be prohibitively expensive
 - Terminal interaction inexpensive, but manual
 - API keys allow for at-scale automation, for order of magnitude cost increase
- Local models require expensive hardware (A100 / H100 GPUs)
- Systems likely be very bespoke, multiple models each doing specific pieces
 - Rare that one model handles all the parsing steps required well.
 - Likely to be a cobbled together system of models which performs best.
- Prompt engineering is critical, and very model dependent
- Validation difficult and a persistent issue. Systems not trustworthy once validated.
- Ensemble / consensus architectures could provide higher confidence.

LLM Model Considerations

- Larger context windows don't necessarily improve performance
 - Individual papers mostly fit in the modern context window sizes
- Page search (RAG) summarization not necessary for individual papers
 - Increases the risk of missing key information contained in the paper
- Science users should not be training their own LLM models.
- Hallucinations remain common and difficult to identify in an automated fashion.
- PDF file encoding not standard in source documents (papers)
 - Creates subtle failures in parsing
- No unified language in science, nuanced domain terminology difficult to parse.
- Should not be used for analysis.
 - Doesn't understand your data, simply performing complex pattern matching.
 - Do not rely on LLMs for math.
- Repeatability questionable
 - Some models give different answers for the same prompt when asked again.