

# AI-Driven Multimodal Validation of Chemical Formulas. Insights from Gemini Thinking Log

Evgeny Markhasin  
Lobachevsky State University of Nizhny Novgorod  
<https://orcid.org/0000-0001-0000-0000>  
[https://\(inkedin.co\)/in/evgenyarkhasin](https://(inkedin.co)/in/evgenyarkhasin)

## Abstract

This work explores prompt engineering strategies for the validation of chemical formulas (as in scholarly manuscripts) using Large Language Models (LLMs). Building upon previous work on LLM-driven peer review, we investigate the efficacy of various prompting techniques from basic series to complex, framework-specific prompts for identifying errors in chemical identifiers. Using a challenging test paper known to contain subtle errors, we observe that simple prompts yield unreliable results. More structured prompts show incremental improvements but still lack robustness. Key insights are derived from analyzing Gemini's thinking logs: this analysis indicates potential challenges related to the model's inherent error-correction tendencies that can mask errors in input, therefore, we then adapt previous demonstrated advanced approaches employing LLM context conditioning combined with explicit instructions for formula validation (analysis of generated results yielding a proof-of-concept prompt for formula validation). This methodology demonstrates improved reliability in identifying errors both in text and formulae. The findings highlight the potential of sophisticated prompting and formula capabilities for chemical entity validation. This work is a supplementary material (for arXiv: 2407.12345v1 [cs.LG] 2024) available at <https://arxiv.org/abs/2407.12345>.

## 1. Introduction

Previous work explored the application of advanced prompt engineering techniques specifically for persistent formulae correction. For the critical analysis of scholarly manuscripts in chemistry, using a specific test paper (test paper) as a challenging case study. This article tends to investigate by focusing on a common yet critical task: the validation of chemical formulas (as within scientific texts). The sample test paper (a 3-page document) included its supporting information. S2/3 serves as a pertinent and challenging test case for this task due to known subtle errors. For instance, page S-B of the test paper; S2 presents the formula for (a) *ferrous ammonium sulfate* as  $(\text{NH}_4)_2\text{Fe}(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$  which is incorrect (its one sulfate group). The correct formula for (a) ferrous ammonium sulfate is  $(\text{NH}_4)_2\text{Fe}(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$  or  $(\text{NH}_4)_2\text{Fe}(\text{SO}_4)_2 \cdot \text{anhydrous}$ . This exploration details the iterative process of developing prompts to reliably identify such chemical formulae (needed in a haystack).

## 2. Approaches to Chemical Formula Validation with LLMs

### 2.1. Direct Prompting Attempts and Initial Observations

Initial efforts involved testing basic prompts with leading LLMs. The initial advanced prompt (Crosby and Ehat 2023) (50% success) prompts such as: Find mistakes in chemical formulas and names. produced highly inconsistent and generally unreliable results (occasionally identifying the target error but lacking systematic accuracy).

1) ore str-ct- red pro) pt +as s- bse5- ent(y deve(oped:

## ## Chemical Formula Extraction and Validation from PDF

Execute the following task step-by-step:

1. Extract each and every chemical formula from the attached PDF.
2. For each extracted formula, extract every directly associated chemical name included in the text, if any.
3. For each extracted formula and associated names, consider if the chemical formula and associated names are correct and flag every formula/names combination that contains any errors.
4. For each flagged item, read the source PDF again and confirm that the item was extracted exactly. In case of any extraction errors, analyze the corrected item and consider if the flag should be removed.
5. Create a Markdown table that should include every flagged formula/names, clear description of any problems, corrected version, and clear reference location of the flagged items.

\*his re4ined pro) pt i) proved the 5- a(ity of the o- tp- t and increased the fre5- ency of identifying the target error in ( 4 )<sub>2</sub> 4 7e) iniNaive1na(ysis. Fo+ever3re(iabi(ity re) ained a signi4icant iss- e.

1n interesting observation arose fro) a s(ight(y) odi4ied version of this pro) pt 7e) iniNaive1na(ysisGeaction. \*he LLM not on(y e, tracted che) ica( for) - (as b- t a(so identi4ied and 4(aggd an i) ba(anced che) ica( reaction sche) e fro) page S-B of the S23i) ) ediate(y fo((o+ ing the prob(e) atic ferro- s a) ) oni- ) s- (fate for) - (a. \*he reaction of De.22/ and hydrogen pero, ide +as 4(aggd d- e to an i) ba(ance in iron ato) s. % De on the (eft3 De on the right/. \*his error3previo- s(y over(ooked by the a- thor d- ring) an- a( revie+ 3+ as correct(y identi4ied and a ba(anced version proposed by the LLM. \*his high(ighted the LLM;s potentia( to - ncover errors beyond the speci4ic target3 even + hen the pro) pt +as foc- sed on 8for) - (as9.

## 2.2. Challenges with LLM Default Behaviors

E, a) ination of 7e) ini;s 8Sho+ thinking9 (ogs. a feat- re providing insight into the ) ode(s processing steps/ for vario- s r- ns revea(ed a consistent pattern. \*he LLM often correct(y e, tracted the target for) - (a. e.g.38 ( 4 )<sub>2</sub> 4: Ferrous ammonium sulfate (Mohr's salt)9/. Fo+ever3in the s- bse5- ent va(idation step3it +o- (d so) eti) es erroneo- s(y) ark the pair as correct3e.g.:

### Identified Chemical Formulas/Names and Initial Validation:

H. Ather (og (ines/H

( 4 )<sub>2</sub> 4: Ferrous ammonium sulfate. Correct.

\*his behavior (ike(y ste) s fro) a core strength of LLMs: their inherent capabi(ity for error correction and - nderstanding intent despite ) inor inacc- racies in the inp- t. Dor instance35- erylmg what is the capital of grate britain? typica((y yie(ds 8London93+ ith the ) isspe((ing of 87reat Oritain9 being a- to) atica((y corrected. < hi(e - s- a((y bene4icia3this defa- (t error to(erance beco) es a hindrance + hen the objective is to *detect* s- ch errors. \*his pheno) enon is ana(ogo- s to the 8inp- t bias9 disc- ssed in the C< C preprint3+ here the LLM;s tendency to accept inp- t infor) ation as fact- a( needs to be active(y co- ntered for critica( eva(- ation.

2n the conte, t of for) - (a va(idation3the LLM;s nat- ra( inc(ination to reconci(e a s(ight(y incorrect for) - (a + ith its correct acco) panying na) e co) p(icates direct 8contrasting9) ethods. as per Coint % in the str- ct- red pro) pt above/. \*o address this prob(e) 3strategies) - st either s- ppress this error-correction tendency in a contro((ed) anner or e) p(oy) ore co) p(e, error detection a(gorith) s that avoid direct co) parison of s- ch 8co) p(e) entary9 yet potentia((y) is) atched entities.

1n ear(y atte) pt at the (atter approach invo(ved a ) - (ti-step generation and co) parison process . see  
81ppendi, . Ehe) ica( Dor) - (a 1na(ysis - 7enerated Dor) - (as and Na) es9 for the pro) pt and an e, a) p(e  
of LLM response/. \*he intent +as for the LLM to:

Extract a chemical formula from text -> Generate a chemical name -> Generate a chemical formula -> Compare the extracted formula with the newly generated formula.

\*his approach3+ hi(e3possib(y3better3sti(( proved - nre(iab(e in practice3a(tho- gh so) e r- ns did yie(d interesting o- tp- ts3inc(- ding references to e, terna( che) ica( databases.

### 3. Advanced Validation using Context Conditioning and Multimodal Analysis

Given the (i) iterations of direct and (ii) (e) generative approaches (3a) more robust strategy + as adopted (3) (leveraging the context conditioning principles outlined in the C < C preprint. \*The *PeerReviewPrompt* detailed in that + work successfully mitigated input bias through comprehensive context setting.

1 ne+ pro) pt3the Ehe) ica(Dor) - (asJa(idationCro) pt3+as deve(oped by adapting core sections fro)  
the *PeerReviewPrompt*:

- K Sections **I-III** . Eore Ablective3Cersona3Eonte, t: Dra) e+ork for Eritica( Gevie+ / +ere (argely retained to estab(ish the ana(ytica() indset and operationa( g- ide(ines.
- K Section **V** . Dina( 2nstr- ctions for 2ninteraction/ +as kept to ens- re consistent LLM behavior.
- K Section **IV.A** . Do- ndationa( Crincip(es L < ork4o+ 1pp(ication/ +as adapted.
- K 1 ne+ **Chemical Identifier Analysis** s- bsection +as introd- ced into Section **IV** speci4ica((y for for) - (a and na) e validation3inc- ding e, p(icit instr- ctions for ) - (ti) oda( ana(ysis of 4ig- res.

\*his C< C-sty(e pro) pt can be s- b) itted to the LLM initia((y . + ith or + itho- t the target C@D/. 1  
s- bse5- ent 5- ery3s- ch as:

## Perform comprehensive Chemical Identifier Analysis

+ith the target C@D attached3then initiates the va(ication +ork4o+.

@e) onstration 12chats - sing this approach + ith 7e) ini . ' Cro . via 7oog(e 12St- dio3+ hich often sho+ s enhanced perfor) ance over the standard s- bscription-based 7e) ini app interface/ 7e) ini1na(ysis and Ehat7C\* C(-s o% Ehat7C\* 1na(ysis sho+ed i) proved rob- stness and re(iabi(ity. \*he c- rrent Ehe) ica(Dor) -(asJa(idationCro) pt instr- cts the LLM to o- tp- t a tab(e of *all* e, tracted che) ica( for) -(as3detai(ing any identi4ied iss- es and providing corrected versions3or a check) ark if no error is fo- nd.

Er-cia((y3this advanced approach consistent(y identi4ied the target error in ( )<sub>4</sub>)<sub>2</sub> 4. D- rther) ore3  
d-e to the e, p(icit instr- ction to perfor) ) - (ti) oda(ana(ysis .speci4ica((y3ana(yzing 4ig- res/3the 7e) ini  
) ode( a(so reprod- cib(y identi4ied another error that had been ) issed in prior ) an- a( revie+s:

2n Dig- re c .page %' of the ) ain te, t of the [test paper](#)/3an NMG spectr-) is presented +ith a (abe(  $\left( \begin{smallmatrix} 3 \\ 3 \end{smallmatrix} \right)_2$  for he, a) ethy(dis(i, o, ane. \*he correct for) - (a is  $\left( \left( \begin{smallmatrix} 3 \\ 3 \end{smallmatrix} \right)_2 \right)$  or  $\left( \begin{smallmatrix} 3 \\ 6 \end{smallmatrix} \right)_2$  . 7e) ini correct(y 4agged the inconsistency in the depicted for) - (a.

< hi(e occasional( fa(se positives + ere observed in the o- tp- ts fro) both ) ode(s3the C< C-based pro) pt + ith ) - (ti) oda( instr- ctions de) onstrated a ) arked i) prove) ent in re(iab(y detecting s- bt(e che) ica( for) - (a errors3inc(- ding those e) bedded + ithin 4ig- res. < hi(e Ehat7C\* o%) ) ode( is advertised as being capab(e of ) - (ti) oda( ana(ysis3it fai(ed to identify the error in the 4ig- re.

## 4. Conclusions

Justifying the idea (for) - (as + within scientific texts - using LLMs presents distinct challenges as illustrated by the - narrative performance of simple or direct strategies and insights from LLM analysis (e.g., these (e.g., for instance) indicate potential challenges related to the - (e.g., inherent error-correction tendencies) which can - ask the very errors targeted for detection and contribute to the observed - narrative ability.

Crucial binary tests demonstrated the efficacy of advanced LLM content conditioning in overcoming this challenge. Specifically, the presented *ChemicalFormulasValidationPrompt* adapted from the previous (y developed C< C-based *PeerReviewPrompt* reproducibly identified the target chemical (for) - (a error - using +o frontier3generally available (LLM) models. D- rather) - (reproducibly) - (ti) oda(analysis + as demonstrated + with the 7e) ini - ' Cro) oda(3+ here a defective chemical (for) - (a + within a 4g- re + as s- ccessf- (y identified by relating it to its name, extracted from the 4g- re caption3and s- bse5- ent(y corrected. Conversely, the same tests failed to elicit effective) - (ti) oda(analysis for this specific task + with the Ehat7C\* C(- s o%) oda(.

\*he demonstrated robustness of content conditioning in managing LLM behaviors - particularly those contributing to input bias and error suppression as evidenced in this study and the previous C< C preprint - suggests potential for broader applications. For instance, similar C< C-informed techniques could prove - self- (in fields such as) medical (123especialy for +ork4o+s involving the critical processing and validation of information from individual (patient records. Another promising application lies in the e, traction and validation of information from - se) i-str- ct- red3- nstr- ct- red3or poorly str- ct- red so- rces3s- ch as those found in pharmaceutical (or technical) documentation3+ here the ability to suppress significant error correction by LLMs may be of paramount importance.

It is important to acknowledge however that the present study itself is based on a limited testing scope. Nonetheless, the presented prompt shows promise as a proof-of-concept for future work should focus on more extensive testing across diverse datasets and further refinement of these C< C-informed methods to enhance accuracy and ensure wider applicability in complex, scientific content analysis and validation.

---

## 5. References

- K **PWP**: Markhasin3E. . 0 ' /. *AI-Driven Scholarly Peer Review via Persistent Workflow Prompting, Meta-Prompting, and Meta-Reasoning*. arXiv preprint. <https://arxiv.org/abs/2407.00000> .
- K **Test paper**: Crasand30.3Le+is31. G3L C(ettner3E. . 0##/. Enrichment of F #! A from \*ap < ater3 Eharacterization of the Enriched < ater3and Properties of Severa( #! A-Labeled Eo) ponds. *Analytical Chemistry*3B%. #/3 %M %\$. <https://doi.org/10.1021/acs.analychem.3c01111> .
- K **ChemicalFormulasValidationPrompt**: The chemical (Dor) - (as Justification Cro) pt. [https://github.com/pchem-g-y/Ehat7C\\*E\\_prompting/blob/main/The\\_chemical\(Dor\)-\(asJustificationCro\)\\_pt/C<C/The\\_chemical\(Dor\)-\(asJustificationCro\)\\_pt.d](https://github.com/pchem-g-y/Ehat7C*E_prompting/blob/main/The_chemical(Dor)-(asJustificationCro)_pt/C<C/The_chemical(Dor)-(asJustificationCro)_pt.d) .
- K **GeminiNaiveAnalysis**: Shared 7e) ini Ehat Log. [https://g.co/gemini/share/a&c=\\$!\\$\\$\\$%&](https://g.co/gemini/share/a&c=$!$$$%&) .
- K **GeminiNaiveAnalysisReaction**: Shared 7e) ini Ehat Log. [https://g.co/gemini/share/!0cd\\_d!b\\_&b](https://g.co/gemini/share/!0cd_d!b_&b) .
- K **GeminiAnalysis**: Shared 7oog(e 12St-dio Ehat Log. [https://aist-dio.google.com/app/prompts/#EUdp=eLstNNp#yOU0\\$G\\$0vaO-hy&EN<#](https://aist-dio.google.com/app/prompts/#EUdp=eLstNNp#yOU0$G$0vaO-hy&EN<#) .
- K **ChatGPTAnalysis**: Shared Ehat7C\* Ehat Log. [https://chatgpt.com/share/6B#f'bSf-%!'c-B00"-b f"- \\$"c!'%!#\\$"](https://chatgpt.com/share/6B#f'bSf-%!'c-B00) .

Appendix. Chemical Formula Analysis - Generated Formulas and Names

Prompt

Execute the following task step-by-step:

1. Extract chemical formulas of each and every chemical species containing at least two elements EXACTLY as they appear in the attached PDF.
2. For each extracted chemical formula generate associated name.
3. Convert each generated name to generated chemical formula.
4. For each generated chemical formula, determine if it matches previously extracted formula.

Response Example (Illustrative)

Chemical Species Analysis from PDF

Here is an analysis of the chemical species found in the provided C@D document. The analysis includes the identification of chemical species, their generated names, and a comparison of the generated names to the original names. The analysis is based on the following criteria:

- Species identified in the provided C@D document.
- Generated names for each species.
- Comparison of the generated names to the original names.

Extracted Chemical Formulas, Generated Names, and Formula Comparison

The following table details the chemical species identified in the provided C@D document, their generated names, and a comparison of the generated names to the original names. The analysis is based on the following criteria:

- Species identified in the provided C@D document.
- Generated names for each species.
- Comparison of the generated names to the original names.

#	Extracted Formula (from PDF)	Generated Chemical Name	Regenerated Chemical Formula (from Name)	Matches Original?	Source(s) for Name/Regenerated Formula
#	F #! A	< ater-#! A	F #! A	Pes	<a href="#">Sig) a-1(drich3</a> <a href="#">Sig) a-1(drich</a>
	E@E(%)	@e- terated ch(orofor)	E@E(%)	Pes	<a href="#">EMOL-E023</a> <a href="#">&lt; ikipedia</a>
%	Na#! AF	Sodi- ) hydro, ide-#! A	Na#! AF .or FNa#! A/	Pes	Na) e ) an- a((y deter) ined d- e to a) big- o- s search: Dor) - (a fro) <a href="#">Ehe) Spider</a>
"	F #! A	Fydrogen pero, ide-#! A	F #! A	Pes	<a href="#">Sig) a-1(drich3</a> <a href="#">Ehe) Spider</a>
'	F A	Fydrogen pero, ide	F A	Pes	<a href="#">Oritannica3US</a> <a href="#">EC13</a> <a href="#">1ccessScience</a>
&	NA	Nitric o, ide .Nitrogen ) ono, ide/	NA	Pes	<a href="#">EMOL-E02</a>
!	F A	< ater .A, idane/	F A	Pes	<a href="#">&lt; ikipedia3</a> <a href="#">C- bEhe)</a>

B	F #BA	< ater-#BA	F #BA	Pes	C- bEhe) 3 N2S* < ebOook
\$	6F CA"	Cotassi-) dihydrogen phosphate	6F CA"	Pes	1) erican Ehe) ica( Society3EMOL- E02
#0	6 FCA"	@ipotassi-) phosphate	6 FCA"	Pes	@r- g0ank3 < ikipedia3 C- bEhe)
##	6E(	Cotassi-) ch(oride	6E(	Pes	OPQ;S3 < ikipedia
#	G- A	G- theni-) .2J/ o, ide	G- A	Pes	< ikipedia
#%	Na2	Sodi-) iodide	Na2	Pes	< ikipedia
#"	F #&A	< ater-#&A	F #&A	Pes	Sig) a-1(drich3 SSF1@E
#'	EF"	Methane	EF"	Pes	EMOL-E023 < ikipedia
#&	FE(	Fydrogen ch(oride	FE(	Pes	< ikipedia3 EMOL-E02
#!	NaAF'	Sodi-) hydro, ide	NaAF'	Pes	< ikipedia3*he Eh(orine 2nstit- te
#B	NaE(	Sodi-) ch(oride	NaE(	Pes	EMOL-E023 *ay(or L Drancis
#\$	Na EA%	Sodi-) carbonate	Na EA%	Pes	Fo) e Science *oo(s3 < ikipedia
0	NaFEA%	Sodi-) bicarbonate	NaFEA%	Pes	< ikipedia3 EOSEA Gesearch Starters
#	Na SA"	Sodi-) s-(fate	Na SA"	Pes	C- bEhe) 3 < ikipedia
	MgSA"	Magnesi-) s-(fate	MgSA"	Pes	D1A
%	ED%	*ri4(- oro) ethy( radica(	ED%	Pes	C- bEhe)
"	NF'	1) ino radica(	NF' .or NF R'	Pes	< ikipedia3 < ikipedia
'	E@%EN	1cetonitri(e-d%	E@%EN	Pes	Merck Mi((ipore3 Sig) a-1(drich
&	EE("	Earbon tetrach(oride	EE("	Pes	C- bEhe)
!	6MnA"	Cotassi-)	6MnA"	Pes	C- bEhe) 3



		per) anganate			@er) Net
B	De.NF"/ SA "	Derro- s a) ) oni-) s- (fate	.NF"/ De.SA"/ or De.NF"/ .SA"/ & F A	No .Arigina( (ike(y a typo or refers to a different species than co) ) on Derro- s 1) ) oni-) S- (fate / Mohr;s Sa(t + hich is a do- b(e sa(t/	EMOL-E03 E1MEA Ehe) ica(s - NA11
\$	6SEN	Cotassi-) thiocyanate	6SEN	Pes	< ikipedia3 Spectr-) Ehe) ica(
%0	De.SEN/%	2ron.222/ thiocyanate	De.SEN/% .or E%DeN%S%/	Pes	C- bEhe) 3 1) erican Ehe) ica( Society
%#	EF%.EF /' AF	#-Fe, ano( .Fe, an-#-o(/	E&F#" A .or EF%.EF /' AF/	Pes	C- bEhe) 3 < ikipedia
%	.EF%/%SiAF	*ri) ethy(si(ano(	.EF%/%SiAF .or E%F#0ASi/	Pes	Sig) a-1(drich3 < ikipedia
%%	..EF%/%Si/ A	Fe, a) ethy(dis(o, ane	..EF%/%Si/ A .or E&F#BASi /	Pes	< ikipedia3 *her) o Disher3 Ereztech
%"	ED%EANF	*ri4(- oroaceta) ide	ED%EANF .or E F D%NA/	Pes	C- bEhe) 3 Disher Scienti4c

**Note on**  $\left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$  4: \*he e, tracted for) - (a  $\left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$  4 does not direct(y correspond to the  
co) ) on co) po- nd Derro- s 1) ) oni-) S- (fate . Mohr;s Sa(t/3+ hich is a do- b(e sa(t + ith the for) - (a  
 $\left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$   $\left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$  or its he, ahydrate  $\left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$   $\left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$  6 2 . \*he e, tracted for) - (a s- ggests a #: # ratio  
of a) ) oni-) to s- (fate3+ hereas Mohr;s sa(t has a : . or #: # after si) p(ifying the s- (fate part if + ritten  
as  $\begin{smallmatrix} 4 \\ 4 \end{smallmatrix} \left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$  4 6 2 /. \*he regenerated for) - (a fro) the na) e 8Derro- s a) ) oni-) s- (fate9  
typica(y refers to the do- b(e sa(t. \*he C@D on page S-B in the s- pporting infor) ation (ists  $\left( \begin{smallmatrix} 4 \\ 2 \end{smallmatrix} \right)_2$  4  
+ hen disc- ssing the 5- antitation of  $\begin{smallmatrix} 2 \\ 2 \end{smallmatrix}$  2. \*his ) ight be a speci4ic notation or a typo in the so- rce  
doc-) ent for + hat is co) ) on(y kno+ n as Mohr;s Sa(t3or a different3(ess co) ) on species.

\*he ana(ysis invo(ved e, tracting for) - (as e, act(y as they appeared in the C@D3- sing + eb searches to  
generate their co) ) on che) ica( na) es3and then - sing those na) es to regenerate che) ica( for) - (as for  
co) parison.