Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

Evidence Contextualization and Counterfactual Attribution for Conversational QA over Heterogeneous Data with RAG Systems

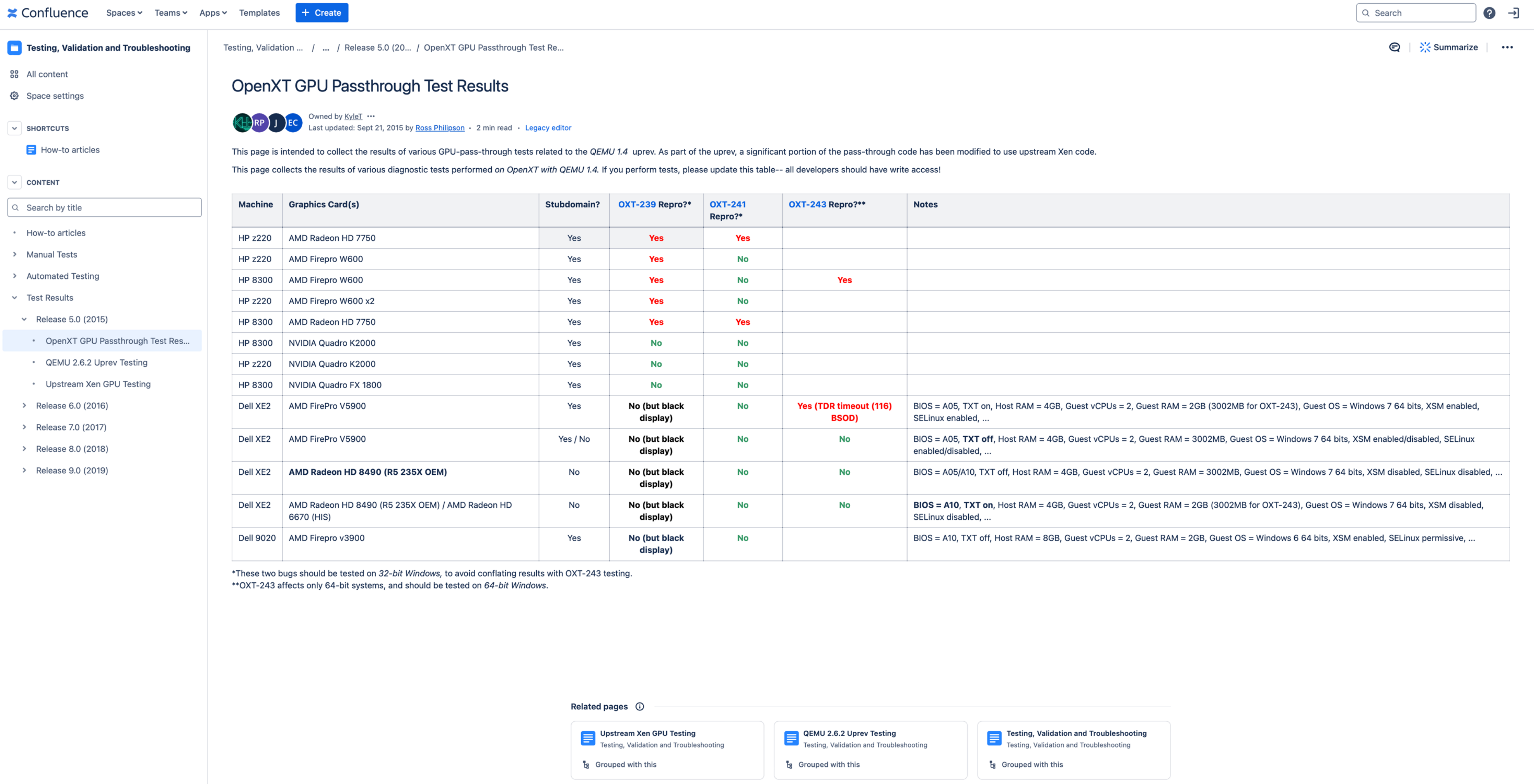
Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech

Use case: Enterprise wiki spaces

Instantiation: Public Confluence pages from developer Atlassian

- **Documents contain heterogeneous elements**
  - Passages, lists, tables
- **Pages are ad hoc HTML**
  - Cannot rely on perfect parsing for tables, lists, passages
- **Niche entities:** Software, configurations, developers
- **Diverse content:** Meeting notes, product descriptions, organizational policies
- **Closely related pages**
  - Software tests for several versions (release candidates 7-9), policies across several years (2015/2016), …
  - Excellent for stress-testing RAG retrieval and generation accuracy
- **"Simulated" enterprise setting**
  - Parametric LLM knowledge is not enough
  - World knowledge not useful for localized questions like "validation team members who attended the oct meeting?"
- Organized into **10 Spaces** with 215 documents: https://openxt.atlassian.net/wiki/spaces

Example page: https://openxt.atlassian.net/wiki/spaces/TEST/pages/6979596/OpenXT+GPU+Passthrough+Test+Results



OpenXT GPU Passthrough Test Results

Owned by KyleT
Last updated: Sept 21, 2015 by Ross Philipson • 2 min read • Legacy editor

This page is intended to collect the results of various GPU-pass-through tests related to the *QEMU 1.4* uprev. As part of the uprev, a significant portion of the pass-through code has been modified to use upstream Xen code.

This page collects the results of various diagnostic tests performed *on OpenXT with QEMU 1.4*. If you perform tests, please update this table-- all developers should have write access!

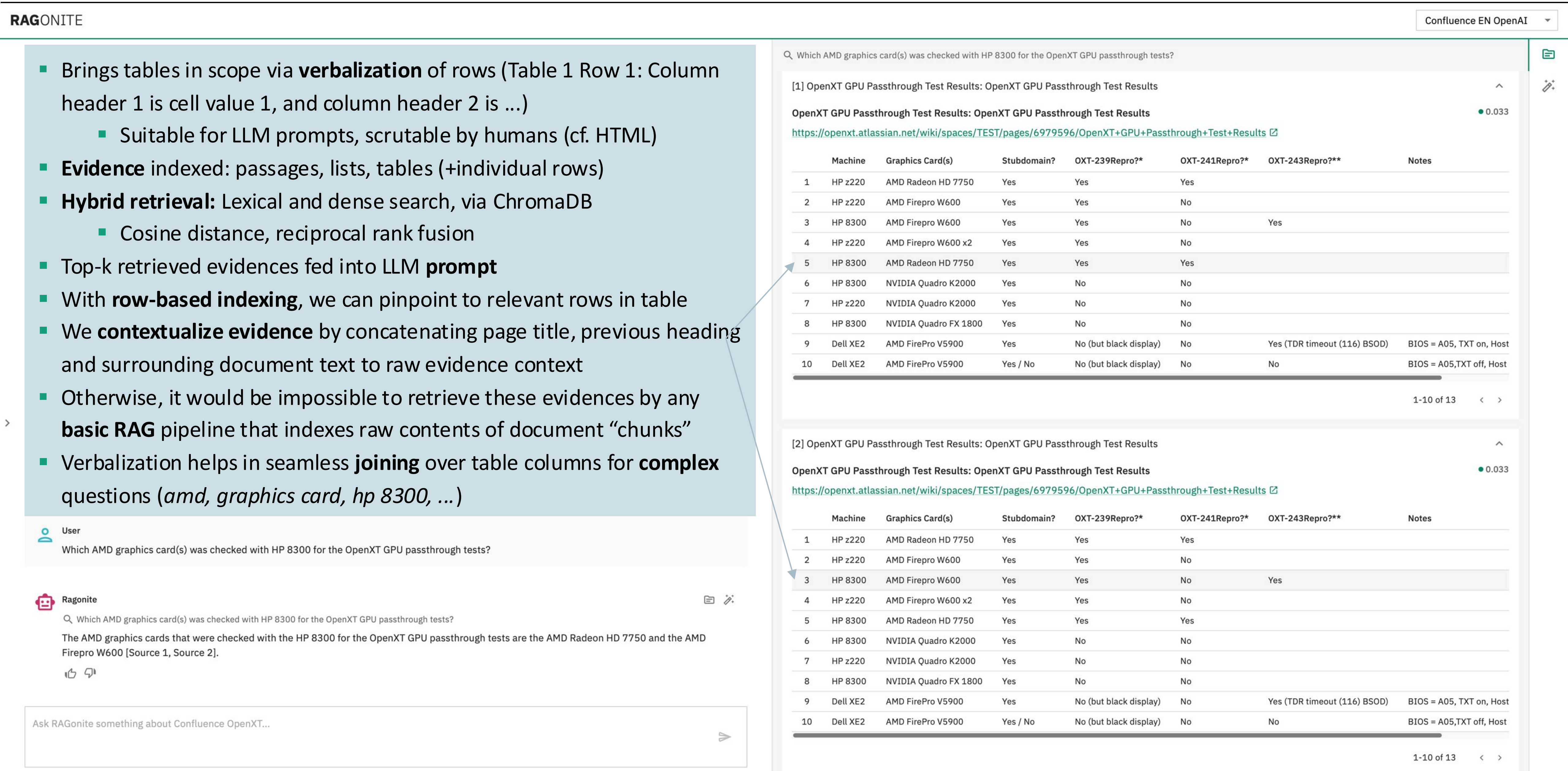
Machine	Graphics Card(s)	Stubdomain?	OXT-239 Repro?*	OXT-241 Repro?*	OXT-243 Repro?**	Notes
HP z220	AMD Radeon HD 7750	Yes	**Yes**	**Yes**		
HP z220	AMD Firepro W600	Yes	**Yes**	**No**		
HP 8300	AMD Firepro W600	Yes	**Yes**	**No**	**Yes**	
HP z220	AMD Firepro W600 x2	Yes	**Yes**	**No**		
HP 8300	AMD Radeon HD 7750	Yes	**Yes**	**Yes**		
HP 8300	NVIDIA Quadro K2000	Yes	**No**	**No**		
HP z220	NVIDIA Quadro K2000	Yes	**No**	**No**		
HP 8300	NVIDIA Quadro FX 1800	Yes	**No**	**No**		
Dell XE2	AMD FirePro V5900	Yes	**No (but black display)**	**No**	**Yes (TDR timeout (116) BSOD)**	BIOS = A05, TXT on, Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS = Windows 7 64 bits, XSM enabled, SELinux enabled, ...
Dell XE2	AMD FirePro V5900	Yes / No	**No (but black display)**	**No**	**No**	BIOS = A05, **TXT off**, Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 3002MB, Guest OS = Windows 7 64 bits, XSM enabled/disabled, SELinux enabled/disabled, ...
Dell XE2	**AMD Radeon HD 8490 (R5 235X OEM)**	No	**No (but black display)**	**No**	**No**	BIOS = A05/A10, TXT off, Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 3002MB, Guest OS = Windows 7 64 bits, XSM disabled, SELinux disabled, ...
Dell XE2	AMD Radeon HD 8490 (R5 235X OEM) / AMD Radeon HD 6670 (HIS)	No	**No (but black display)**	**No**	**No**	**BIOS = A10**, **TXT on**, Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS = Windows 7 64 bits, XSM disabled, SELinux disabled, ...
Dell 9020	AMD Firepro v3900	Yes	**No (but black display)**	**No**		BIOS = A10, TXT off, Host RAM = 8GB, Guest vCPUs = 2, Guest RAM = 2GB, Guest OS = Windows 6 64 bits, XSM enabled, SELinux permissive, ...

*These two bugs should be tested on *32-bit Windows,* to avoid conflating results with OXT-243 testing.
**OXT-243 affects only 64-bit systems, and should be tested on *64-bit Windows*.

Related pages
Upstream Xen GPU Testing
Testing, Validation and Troubleshooting
Grouped with this
QEMU 2.6.2 Uprev Testing
Testing, Validation and Troubleshooting
Grouped with this
Testing, Validation and Troubleshooting
Testing, Validation and Troubleshooting
Grouped with this

Example question 1: Complex questions on structured evidence supported

Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?



- Brings tables in scope via **verbalization** of rows (Table 1 Row 1: Column header 1 is cell value 1, and column header 2 is …)
  - Suitable for LLM prompts, scrutable by humans (cf. HTML)
- **Evidence** indexed: passages, lists, tables (+individual rows)
- **Hybrid retrieval:** Lexical and dense search, via ChromaDB
  - Cosine distance, reciprocal rank fusion
- Top-k retrieved evidences fed into LLM **prompt**
- With **row-based indexing**, we can pinpoint to relevant rows in table
- We **contextualize evidence** by concatenating page title, previous heading and surrounding document text to raw evidence context
- Otherwise, it would be impossible to retrieve these evidences by any **basic RAG** pipeline that indexes raw contents of document "chunks"
- Verbalization helps in seamless **joining** over table columns for **complex** questions (*amd, graphics card, hp 8300, …*)

User

Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

Ragonite

Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

The AMD graphics cards that were checked with the HP 8300 for the OpenXT GPU passthrough tests are the AMD Radeon HD 7750 and the AMD Firepro W600 [Source 1, Source 2].

Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

[1] OpenXT GPU Passthrough Test Results: OpenXT GPU Passthrough Test Results

OpenXT GPU Passthrough Test Results: OpenXT GPU Passthrough Test Results ● 0.033

https://openxt.atlassian.net/wiki/spaces/TEST/pages/6979596/OpenXT+GPU+Passthrough+Test+Results

	Machine	Graphics Card(s)	Stubdomain?	OXT-239Repro?*	OXT-241Repro?*	OXT-243Repro?**	Notes
1	HP z220	AMD Radeon HD 7750	Yes	Yes	Yes		
2	HP z220	AMD Firepro W600	Yes	Yes	No		
3	HP 8300	AMD Firepro W600	Yes	Yes	No	Yes	
4	HP z220	AMD Firepro W600 x2	Yes	Yes	No		
5	HP 8300	AMD Radeon HD 7750	Yes	Yes	Yes		
6	HP 8300	NVIDIA Quadro K2000	Yes	No	No		
7	HP z220	NVIDIA Quadro K2000	Yes	No	No		
8	HP 8300	NVIDIA Quadro FX 1800	Yes	No	No		
9	Dell XE2	AMD FirePro V5900	Yes	No (but black display)	No	Yes (TDR timeout (116) BSOD)	BIOS = A05, TXT on, Host
10	Dell XE2	AMD FirePro V5900	Yes / No	No (but black display)	No	No	BIOS = A05,TXT off, Host

1-10 of 13

[2] OpenXT GPU Passthrough Test Results: OpenXT GPU Passthrough Test Results

OpenXT GPU Passthrough Test Results: OpenXT GPU Passthrough Test Results ● 0.033

https://openxt.atlassian.net/wiki/spaces/TEST/pages/6979596/OpenXT+GPU+Passthrough+Test+Results

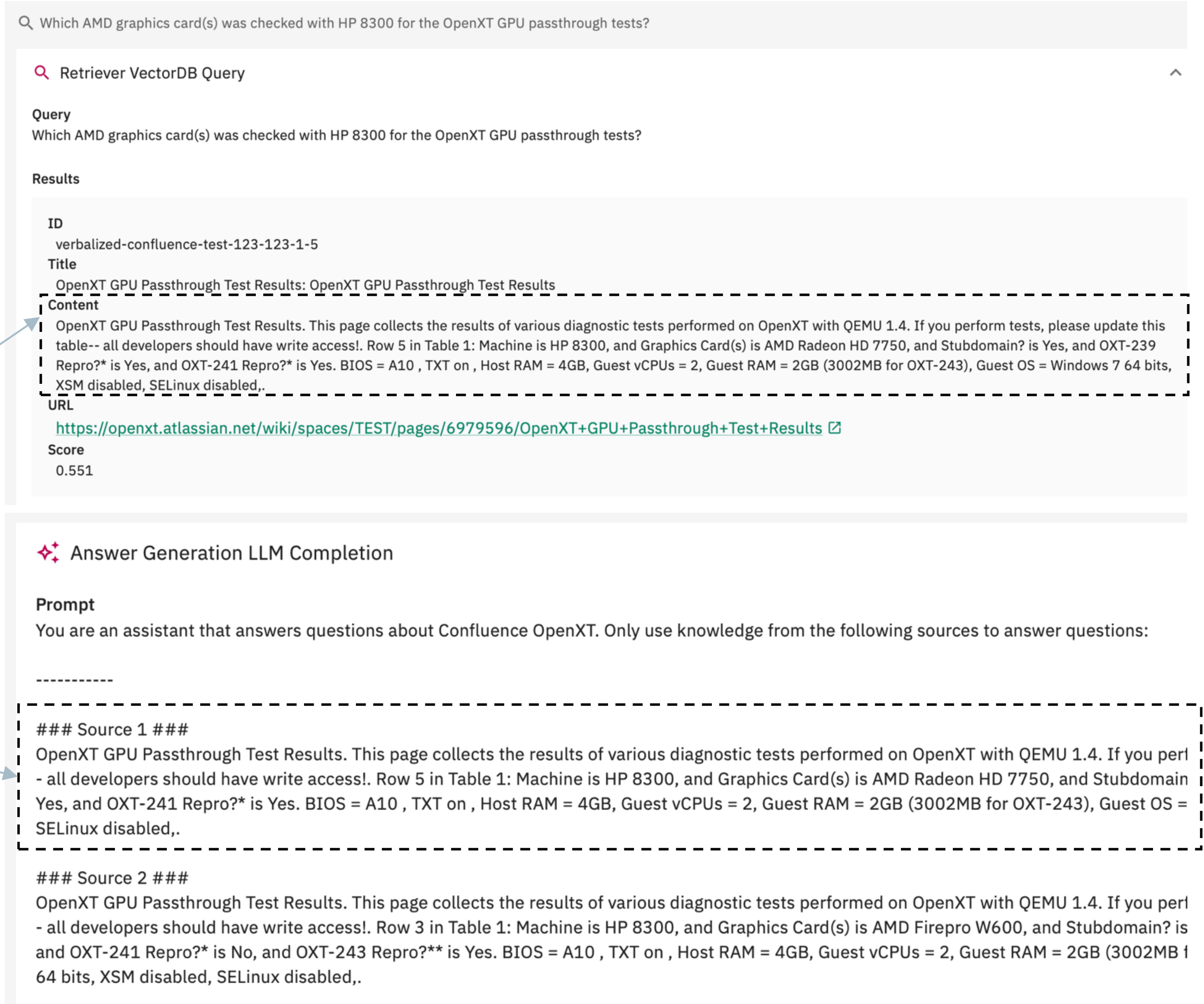
	Machine	Graphics Card(s)	Stubdomain?	OXT-239Repro?*	OXT-241Repro?*	OXT-243Repro?**	Notes
1	HP z220	AMD Radeon HD 7750	Yes	Yes	Yes		
2	HP z220	AMD Firepro W600	Yes	Yes	No		
3	HP 8300	AMD Firepro W600	Yes	Yes	No	Yes	
4	HP z220	AMD Firepro W600 x2	Yes	Yes	No		
5	HP 8300	AMD Radeon HD 7750	Yes	Yes	Yes		
6	HP 8300	NVIDIA Quadro K2000	Yes	No	No		
7	HP z220	NVIDIA Quadro K2000	Yes	No	No		
8	HP 8300	NVIDIA Quadro FX 1800	Yes	No	No		
9	Dell XE2	AMD FirePro V5900	Yes	No (but black display)	No	Yes (TDR timeout (116) BSOD)	BIOS = A05, TXT on, Host
10	Dell XE2	AMD FirePro V5900	Yes / No	No (but black display)	No	No	BIOS = A05,TXT off, Host

1-10 of 13

Example question 1: Complex questions on structured evidence supported

Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

- **Interpretable pipeline**
  - Every intermediate result is transparent and interpretable
  - Retriever queries, evidence details (scores, content, URL)
  - LLM prompts and instructions
- **Behind-the-scenes**
  - Proof of evidence contextualization in dashed box
  - Content starts at ”Row 5 in Table 1: Machine is HP 8300, and ...”
  - Without prepending page title (*openxt gpu passthrough tests*) evidence could not have been retrieved
  - LLM answer could not have been generated



Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

Retriever VectorDB Query

Query
Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

Results

ID
verbalized-confluence-test-123-123-1-5
Title
OpenXT GPU Passthrough Test Results: OpenXT GPU Passthrough Test Results
Content
OpenXT GPU Passthrough Test Results. This page collects the results of various diagnostic tests performed on OpenXT with QEMU 1.4. If you perform tests, please update this table-- all developers should have write access!. Row 5 in Table 1: Machine is HP 8300, and Graphics Card(s) is AMD Radeon HD 7750, and Stubdomain? is Yes, and OXT-239 Repro?* is Yes, and OXT-241 Repro?* is Yes. BIOS = A10 , TXT on , Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS = Windows 7 64 bits, XSM disabled, SELinux disabled,.
URL
https://openxt.atlassian.net/wiki/spaces/TEST/pages/6979596/OpenXT+GPU+Passthrough+Test+Results
Score
0.551

Answer Generation LLM Completion

Prompt
You are an assistant that answers questions about Confluence OpenXT. Only use knowledge from the following sources to answer questions:

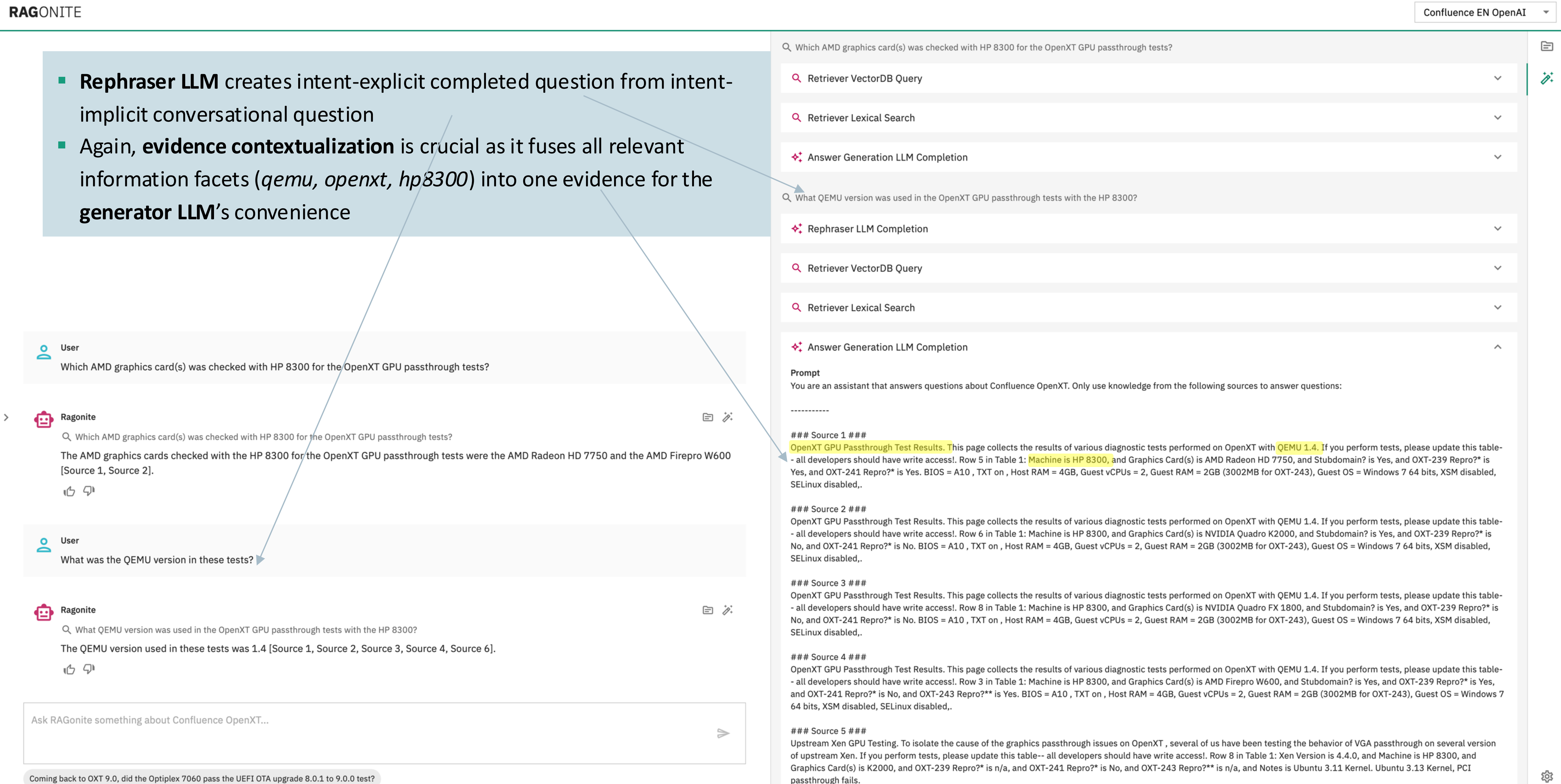
Source 1
OpenXT GPU Passthrough Test Results. This page collects the results of various diagnostic tests performed on OpenXT with QEMU 1.4. If you perf
- all developers should have write access!. Row 5 in Table 1: Machine is HP 8300, and Graphics Card(s) is AMD Radeon HD 7750, and Stubdomain
Yes, and OXT-241 Repro?* is Yes. BIOS = A10 , TXT on , Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS =
SELinux disabled,.

Source 2
OpenXT GPU Passthrough Test Results. This page collects the results of various diagnostic tests performed on OpenXT with QEMU 1.4. If you perf
- all developers should have write access!. Row 3 in Table 1: Machine is HP 8300, and Graphics Card(s) is AMD Firepro W600, and Stubdomain? is
and OXT-241 Repro?* is No, and OXT-243 Repro?** is Yes. BIOS = A10 , TXT on , Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB f
64 bits, XSM disabled, SELinux disabled,.

Example question 2: Conversational questions with implicit context supported

What was the QEMU version in these tests? (need to resolve "these tests" to previous configurations)

- **Rephraser LLM** creates intent-explicit completed question from intent-implicit conversational question
- Again, **evidence contextualization** is crucial as it fuses all relevant information facets (*qemu, openxt, hp8300*) into one evidence for the **generator LLM**'s convenience



User
Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

Ragonite
Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?
The AMD graphics cards checked with the HP 8300 for the OpenXT GPU passthrough tests were the AMD Radeon HD 7750 and the AMD Firepro W600 [Source 1, Source 2].

User
What was the QEMU version in these tests?

Ragonite
What QEMU version was used in the OpenXT GPU passthrough tests with the HP 8300?
The QEMU version used in these tests was 1.4 [Source 1, Source 2, Source 3, Source 4, Source 6].

Ask RAGonite something about Confluence OpenXT...
Coming back to OXT 9.0, did the Optiplex 7060 pass the UEFI OTA upgrade 8.0.1 to 9.0.0 test?

Which AMD graphics card(s) was checked with HP 8300 for the OpenXT GPU passthrough tests?

What QEMU version was used in the OpenXT GPU passthrough tests with the HP 8300?

Prompt
You are an assistant that answers questions about Confluence OpenXT. Only use knowledge from the following sources to answer questions:

Source 1
OpenXT GPU Passthrough Test Results. This page collects the results of various diagnostic tests performed on OpenXT with QEMU 1.4. If you perform tests, please update this table-- all developers should have write access!. Row 5 in Table 1: Machine is HP 8300, and Graphics Card(s) is AMD Radeon HD 7750, and Stubdomain? is Yes, and OXT-239 Repro?* is Yes, and OXT-241 Repro?* is Yes. BIOS = A10 , TXT on , Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS = Windows 7 64 bits, XSM disabled, SELinux disabled,.

Source 2
OpenXT GPU Passthrough Test Results. This page collects the results of various diagnostic tests performed on OpenXT with QEMU 1.4. If you perform tests, please update this table-- all developers should have write access!. Row 6 in Table 1: Machine is HP 8300, and Graphics Card(s) is NVIDIA Quadro K2000, and Stubdomain? is Yes, and OXT-239 Repro?* is No, and OXT-241 Repro?* is No. BIOS = A10 , TXT on , Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS = Windows 7 64 bits, XSM disabled, SELinux disabled,.

Source 3
OpenXT GPU Passthrough Test Results. This page collects the results of various diagnostic tests performed on OpenXT with QEMU 1.4. If you perform tests, please update this table-- all developers should have write access!. Row 8 in Table 1: Machine is HP 8300, and Graphics Card(s) is NVIDIA Quadro FX 1800, and Stubdomain? is Yes, and OXT-239 Repro?* is No, and OXT-241 Repro?* is No. BIOS = A10 , TXT on , Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS = Windows 7 64 bits, XSM disabled, SELinux disabled,.

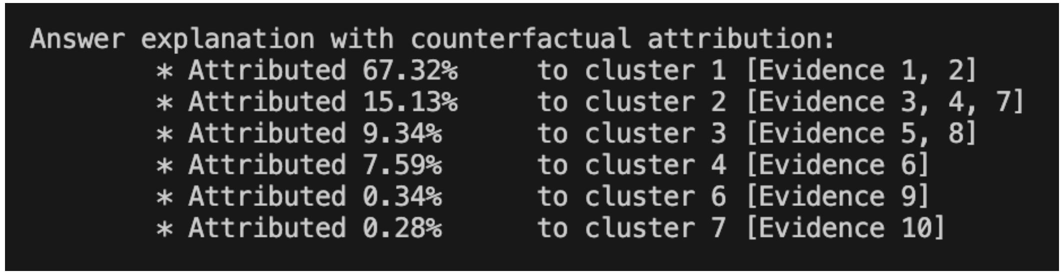
Source 4
OpenXT GPU Passthrough Test Results. This page collects the results of various diagnostic tests performed on OpenXT with QEMU 1.4. If you perform tests, please update this table-- all developers should have write access!. Row 3 in Table 1: Machine is HP 8300, and Graphics Card(s) is AMD Firepro W600, and Stubdomain? is Yes, and OXT-239 Repro?* is Yes, and OXT-241 Repro?* is No, and OXT-243 Repro?** is Yes. BIOS = A10 , TXT on , Host RAM = 4GB, Guest vCPUs = 2, Guest RAM = 2GB (3002MB for OXT-243), Guest OS = Windows 7 64 bits, XSM disabled, SELinux disabled,.

Source 5
Upstream Xen GPU Testing. To isolate the cause of the graphics passthrough issues on OpenXT , several of us have been testing the behavior of VGA passthrough on several version of upstream Xen. If you perform tests, please update this table-- all developers should have write access!. Row 8 in Table 1: Xen Version is 4.4.0, and Machine is HP 8300, and Graphics Card(s) is K2000, and OXT-239 Repro?* is n/a, and OXT-241 Repro?* is No, and OXT-243 Repro?** is n/a, and Notes is Ubuntu 3.11 Kernel. Ubuntu 3.13 Kernel, PCI passthrough fails.

Answer explanation via counterfactual attribution

Counterfactual evidence leads to stronger causal inference than statistical association

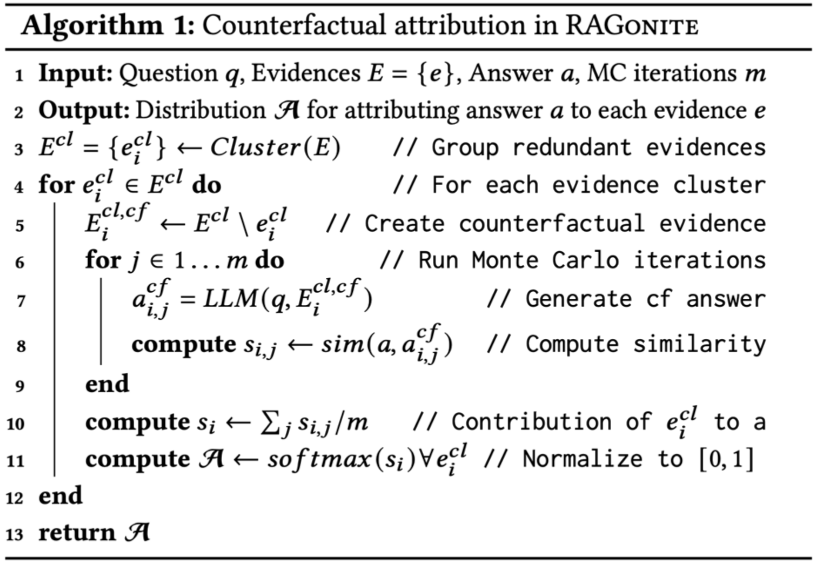
- **Attribution is a popular mechanism of explaining RAG answers**
  - Each evidence gets contribution score
  - Typically based on evidence-answer similarity (statistical association)
- **We propose using counterfactual evidence**
  - RAGonite regenerates response after removing each evidence
  - Similarity between original and counterfactual answer => contribution of evidence removed
  - High similarity implies low contribution, and vice versa
  - Softmax over raw similarity distribution produces final attribution distribution
- **Redundant evidences are a confounding factor**
  - Removal of evidence may lead to no change in answer in presence of equivalent replacement
  - Evidences need to be clustered before removal
- **Fine-tuning similarity functions** can lead to more discriminative distributions
- **Parallelization over evidence clusters** keeps inference times under 2 seconds



```
Answer explanation with counterfactual attribution:
        * Attributed 67.32%     to cluster 1 [Evidence 1, 2]
        * Attributed 15.13%     to cluster 2 [Evidence 3, 4, 7]
        * Attributed 9.34%      to cluster 3 [Evidence 5, 8]
        * Attributed 7.59%      to cluster 4 [Evidence 6]
        * Attributed 0.34%      to cluster 6 [Evidence 9]
        * Attributed 0.28%      to cluster 7 [Evidence 10]
```

Sample output

Formal algorithm

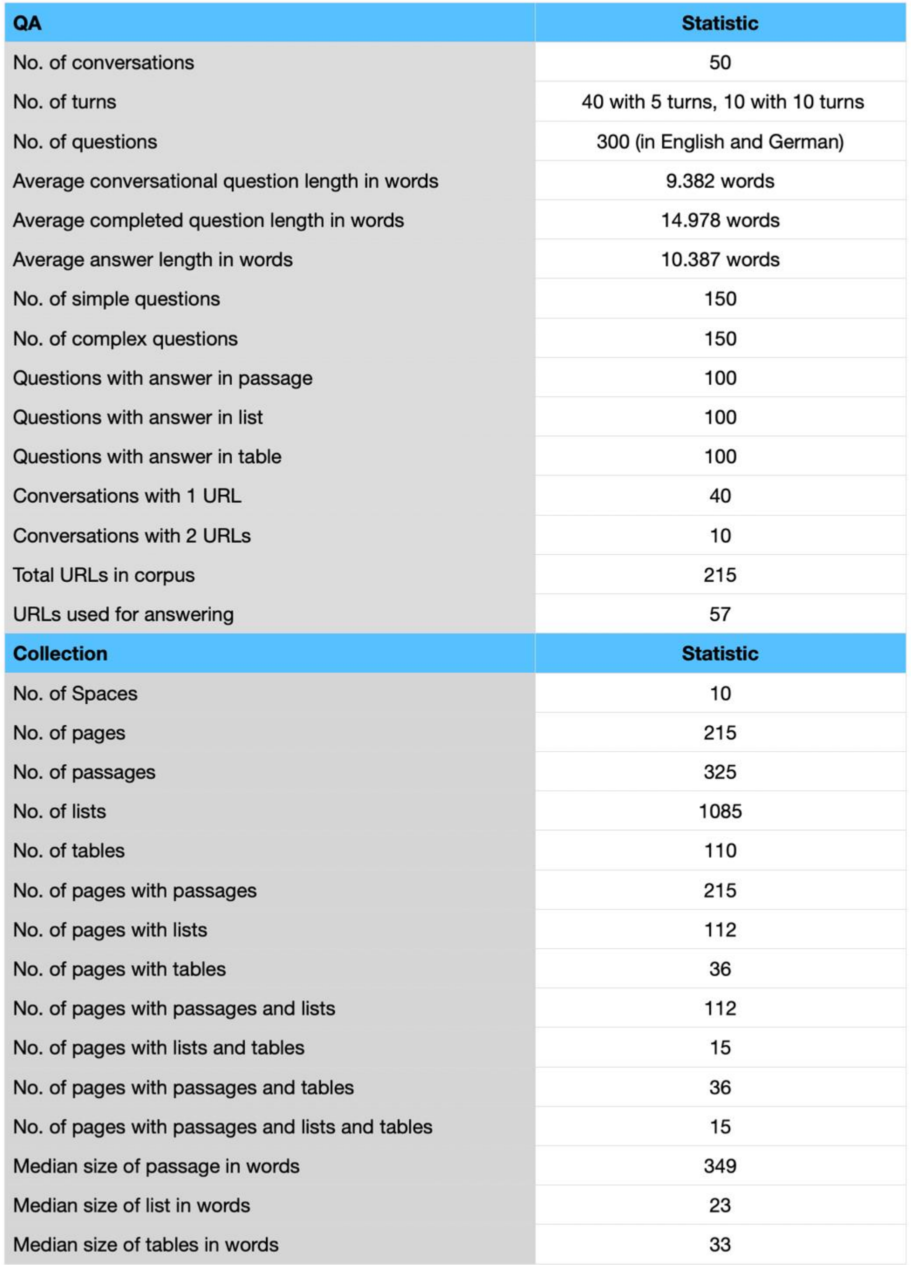


```
Algorithm 1: Counterfactual attribution in RAGONITE
Input: Question $q$, Evidences $E = \{e\}$, Answer $a$, MC iterations $m$
Output: Distribution $\mathcal{A}$ for attributing answer $a$ to each evidence $e$
$E^{cl} = \{e_i^{cl}\} \leftarrow Cluster(E)$      // Group redundant evidences
for $e_i^{cl} \in E^{cl}$ do                        // For each evidence cluster
    $E_i^{cl,cf} \leftarrow E^{cl} \setminus e_i^{cl}$   // Create counterfactual evidence
    for $j \in 1 \ldots m$ do                       // Run Monte Carlo iterations
        $a_{i,j}^{cf} = LLM(q, E_i^{cl,cf})$            // Generate cf answer
        compute $s_{i,j} \leftarrow sim(a, a_{i,j}^{cf})$  // Compute similarity
    end
    compute $s_i \leftarrow \sum_j s_{i,j}/m$        // Contribution of $e_i^{cl}$ to a
    compute $\mathcal{A} \leftarrow softmax(s_i) \forall e_i^{cl}$  // Normalize to $[0,1]$
end
return $\mathcal{A}$
```

The ConfQuestions benchmark

Evaluates question rephraser, evidence retriever, answer generator

- **ConfQuestions is suitable for**
  - Conversational question answering
  - Complex question answering
  - RAG systems
  - Heterogeneous documents (with passages, lists, tables and combinations)
  - Enterprise wiki spaces
- Existing QA benchmarks do not possess all of above desiderata
- **ConfQuestions contains**
  - 300 conversational questions in all (human generated) organized into 50 conversations
  - Corresponding completed questions (human generated)
  - Gold URL and answers: entities, phrases, lists, passages (human generated)
  - Each question in English and German (human translated)
  - Metadata
    - Answer-source (passage/list/table), Complexity type (simple/complex)



QA	Statistic
No. of conversations	50
No. of turns	40 with 5 turns, 10 with 10 turns
No. of questions	300 (in English and German)
Average conversational question length in words	9.382 words
Average completed question length in words	14.978 words
Average answer length in words	10.387 words
No. of simple questions	150
No. of complex questions	150
Questions with answer in passage	100
Questions with answer in list	100
Questions with answer in table	100
Conversations with 1 URL	40
Conversations with 2 URLs	10
Total URLs in corpus	215
URLs used for answering	57
Collection	**Statistic**
No. of Spaces	10
No. of pages	215
No. of passages	325
No. of lists	1085
No. of tables	110
No. of pages with passages	215
No. of pages with lists	112
No. of pages with tables	36
No. of pages with passages and lists	112
No. of pages with lists and tables	15
No. of pages with passages and tables	36
No. of pages with passages and lists and tables	15
Median size of passage in words	349
Median size of list in words	23
Median size of tables in words	33

Thank you!

Looking forward to your feedback!