ALEPH ALPHA

LUMINOUS
Performance Benchmarks

We present our benchmark results for Aleph Alpha's Large Language Models: **luminous-base**, **luminous-extended** and **luminous-supreme** available on the Completion Playground and API client. Luminous models follow a decoder-only autoregressive architecture with use of rotary positional embeddings. Our models are trained on a curated multilingual corpus containing sources in English, German, French, Italian, and Spanish on ~400B to ~588B language tokens for the smallest and largest models, respectively.

February 2023

www.aleph-alpha.com/research

Table of contents

1. Evaluation framework

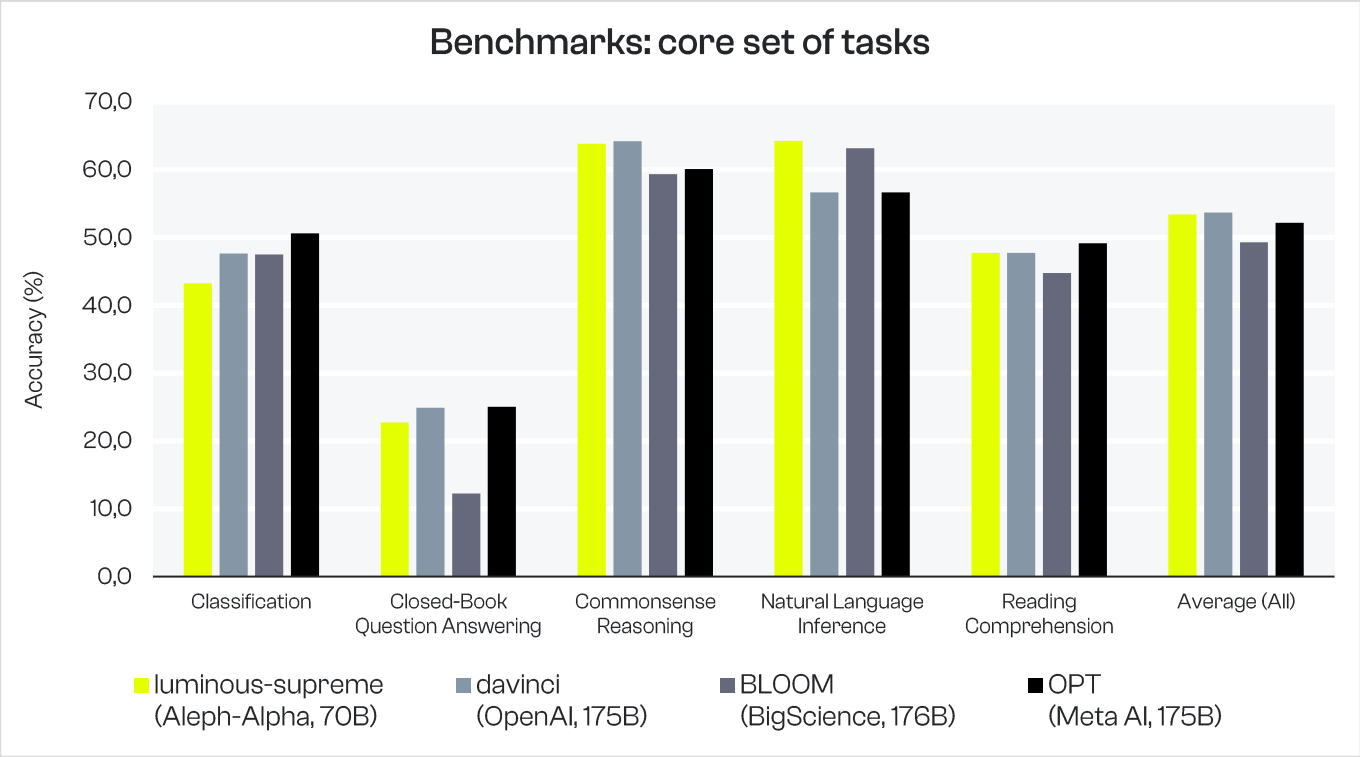
While the Aleph Alpha Playground and API also provide Question-Answering, Embedding, and Summarization endpoints along with multimodal capabilities, here we focus on evaluating the Large Language Model's text-based completions. For that, EleutherAI's Evaluation Harness (lm-eval) package is used.

Completion correctness is measured with the soft accuracy metric (acc) when possible; here the log-likelihood probability is measured for each of the possible multiple-choice completion options, and the one with the highest probability is selected to determine the accuracy of the prediction compared to the ground-truth option. For generative tasks, an exact match accuracy metric (exact) is computed by checking if model completion exactly matches the expected output. The tasks evaluated with the exact metric are the following: squad2, triviaqs, webqs.

Note that when comparing to other models, we only do so for results produced with this common benchmarking setup, as evaluation results can deviate from published ones for some tasks due to prompt formulation, checkpoint formats, data splits used, etc.

2. Model performance on core tasks

Firstly, we show results for our current best model (**luminous-supreme** with 70B parameters) on a set of core 16 tasks and compared to BigScience **BLOOM** (176B) and Meta AI **OPT** (175B) benchmark results produced with **lm-eval** (see link here). OpenAI **davinci** (175B) is evaluated by ourselves using the same setup as for our Aleph Alpha **Luminous** models. We note competitive accuracy results when averaged across all 16 tasks, especially given our smaller architecture in comparison to the other models.



The benchmarked tasks cover five groups:

- Classification (**wic**),
- Closed-Book Question Answering (**triviaqa**, **webqs**),
- Common-sense Reasoning (**arc_challenge**, **arc_easy**, **copa**, **hellaswag**, **openbookqa**, **piqa**, **winogrande**, **wsc**),
- Natural Language Inference (**rte**),
- Reading Comprehension (**boolq**, **lambada**, **multirc**, **race**).

Example prompts for the evaluated tasks, and detailed information on performance are provided in the supplementary material section in chapter 6.

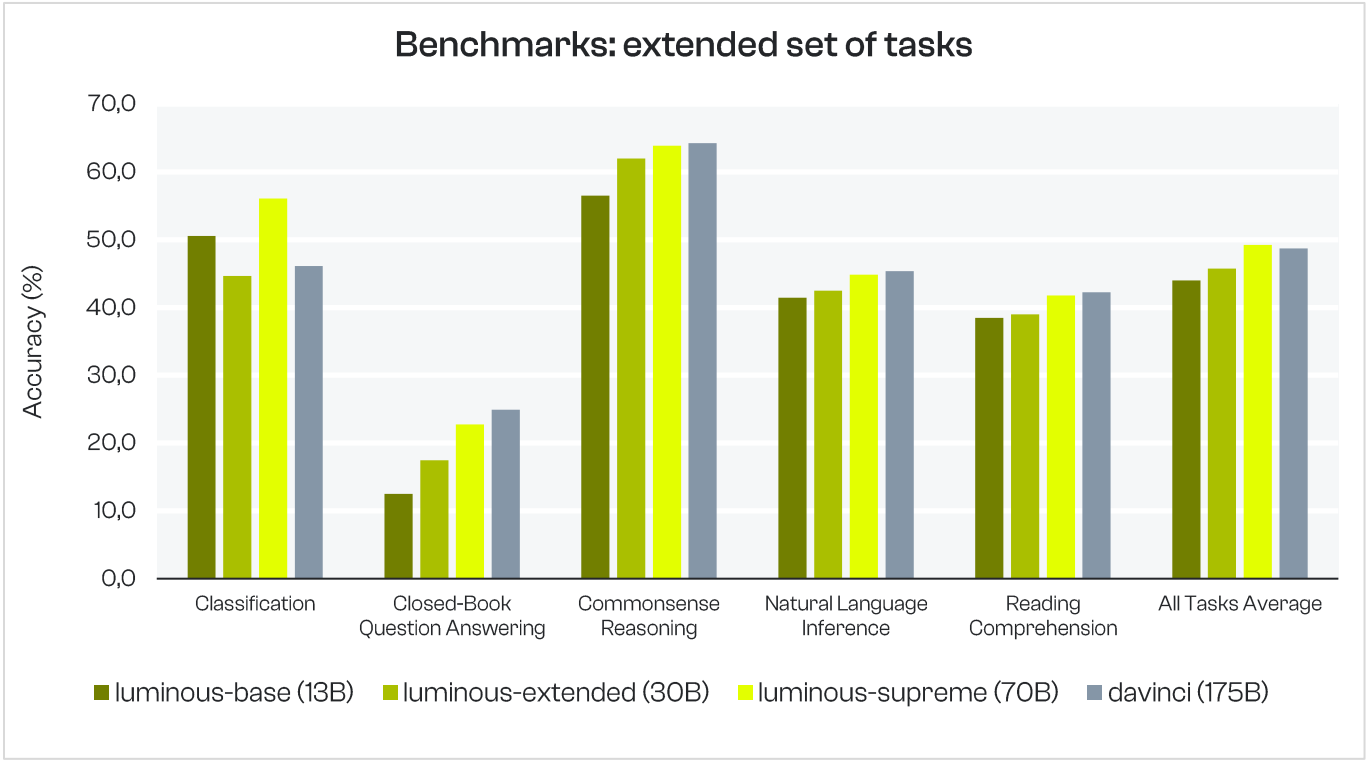
3. Model performance on an extended set of tasks

We also release additional benchmark results with 28 additional tasks, extending the evaluation set to a total of 44 tasks. We add to our comparison our two alternative smaller models, **luminous-base** and **luminous-extended** with 13B and 30B parameters, respectively.

The additional tasks are:

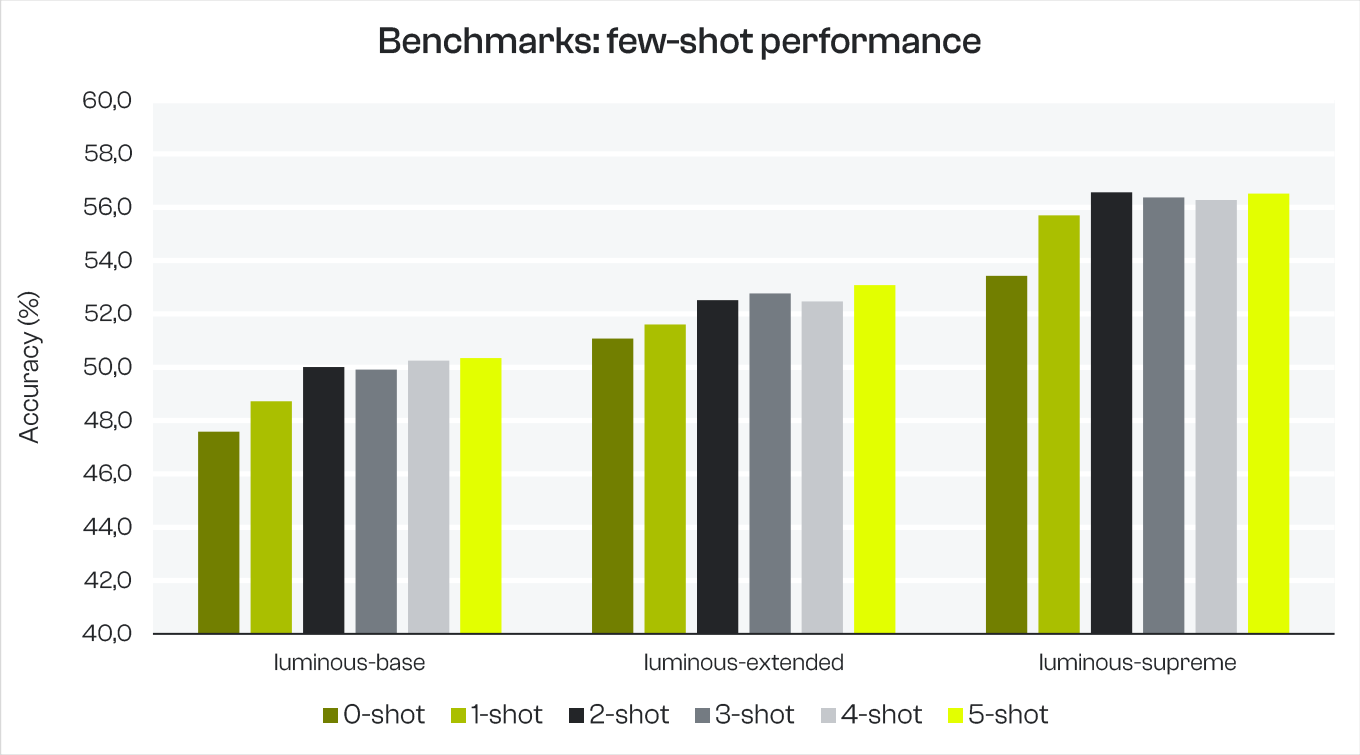
- Classification (**mrpc**, **sst**),
- Natural Language Inference (**anli_r1**, **anli_r2**, **anli_r3**, **cb**, **mnli**, **mnli_mismatched**, **qnli**, **wnli**),
- Reading Comprehension (**squad2**, **race_mid**).

We can see from the graph below that performance improvements are seen for all task types evaluated in the results, as model size increases. This trend is in line with empirical scaling law observations as seen in prior work. We expect further improvements with our scaled-up **luminous-world** model (coming soon).



4. Few-shot prompting

The table below provides the benchmarking results with few-shot prompting for the 16 core tasks listed above. Zero to five examples (separated by two new lines in **lm-eval**) are concatenated in the prompt as input for the completion.



Few-shot prompting helps to boost performance in completion tasks for our Luminous models (6% on **luminous-supreme** between 0-shot and 5-shots). Depending on the end-user needs, this also allows for faster and cheaper inference using smaller models without compromising on accuracy: for example, providing 5-shot examples on **luminous-extended** can enable performance that is almost as good as 0-shot **luminous-supreme** on average.

While the table compares likelihood-based accuracies, we observed an even larger impact on exact-match completion tasks, where the few-shot examples help boost the performance by providing the model with examples of what type of answers are expected.

In addition, we did an ablation study where **luminous-supreme** was evaluated with 1-shot examples again but with "\n###\n" as a separator (which is usually used in our examples in the Playground) rather than the **lm-eval** standard new lines "\n\n". Using both few-shot separators gives very close average accuracy scores (55.3% with "\n###\n" and 55.7% with "\n\n") and scores for individual tasks are within 5% of each other.

5. Supplementary materials

Benchmark results on the core set of tasks

task & category		metric	luminous-base (Aleph-Alpha, 13B)	luminous-extended (Aleph-Alpha, 30B)	luminous-supreme (Aleph-Alpha, 70B)	davinci (OpenAI, 175B)	BLOOM (BigScience, 176B)	OPT (Meta AI, 175B)
arc_challenge	CR	acc	37,0	40,7	44,2	43,2	41,1	41,2
arc_easy	CR	acc	70,3	75,0	76,9	76,5	72,6	75,1
boolq	RC	acc	68,3	69,1	74,3	73,4	70,4	80,2
copa	CR	acc	83,0	84,0	90,0	89,0	87,0	84,0
hellaswag	CR	acc	53,3	56,9	59,2	59,1	53,5	59,2
lambada	RC	acc	70,2	72,5	74,0	75,1	67,2	74,7
multirc	RC	acc	1,7	1,6	1,9	4,0	2,4	1,6
openbookqa	CR	acc	27,8	29,4	33,0	33,6	31,2	32,2
piqa	CR	acc	77,3	78,8	79,9	79,1	78,1	79,1
race	RC	acc	37,2	38,6	40,8	38,6	39,0	40,2
rte	NLI	acc	57,0	57,4	64,3	56,7	63,2	56,7
triviaqa	QA	exact	20,5	28,7	37,7	40,9	18,3	34,2
webqs	QA	exact	4,5	6,2	7,8	8,9	6,2	15,9
wic	CL	acc	49,8	47,5	43,3	47,6	47,5	50,6
winogrande	CR	acc	64,6	67,3	70,0	69,9	71,0	73,6
wsc	CR	acc	38,5	63,5	57,7	63,5	40,4	36,5
Classification			49,8	47,5	43,3	47,6	47,5	50,6
Closed-Book Question Answering			12,5	17,4	22,7	24,9	12,2	25,0
Commonsense Reasoning			56,5	62,0	63,9	64,2	59,4	60,1
Natural Language Inference			57,0	57,4	64,3	56,7	63,2	56,7
Reading Comprehension			44,4	45,4	47,7	47,8	44,8	49,2
Average (All)			47,6	51,1	53,4	53,7	49,3	52,2

Benchmark results on the extended set of tasks

task & category		metric	luminous-base (13B)	luminous-extended (30B)	luminous-supreme (70B)	davinci (175B)
anli_r1	NLI	acc	33,3	30,8	33,7	36,1
anli_r2	NLI	acc	33,5	35,5	34,1	37,5
anli_r3	NLI	acc	33,6	36,8	36,1	36,8
cb	NLI	acc	41,1	30,4	35,7	41,1
squad2	RC	exact	9,9	10,7	18,9	21,1
mnli	NLI	acc	36,0	42,0	45,9	39,6
mnli_mismatched	NLI	acc	35,7	43,7	47,1	41,0
mrpc	CL	acc	47,8	36,5	65,4	39,5
qnli	NLI	acc	50,7	53,6	50,5	52,0
race_mid	RC	acc	43,4	41,4	40,9	41,4
sst	CL	acc	54,0	49,9	59,5	51,3
wnli	NLI	acc	52,1	52,1	56,3	67,6
Classification			50,6	44,6	56,1	46,1
Closed-Book Question Answering			12,5	17,4	22,7	24,9
Commonsense Reasoning			56,5	62,0	63,9	64,2
Natural Language Inference			41,4	42,5	44,9	45,4
Reading Comprehension			38,5	39,0	41,8	42,3
All Tasks Average			44,0	45,7	49,3	48,7

Benchmark with few-shot prompts

task & category		metric	luminous-base						luminous-extended						luminous-supreme					
			0-shot	1-shot	2-shot	3-shot	4-shot	5-shot	0-shot	1-shot	2-shot	3-shot	4-shot	5-shot	0-shot	1-shot	2-shot	3-shot	4-shot	5-shot
arc_challenge	CR	acc	37,0	37,5	39,2	39,6	38,8	39,1	40,7	42,7	44,5	44,2	44,1	44,8	44,2	47,8	49,7	49,7	50,0	50,3
arc_easy	CR	acc	70,3	72,4	72,6	72,8	73,2	73,4	75,0	75,9	76,9	77,1	76,8	76,6	76,9	79,0	79,5	79,9	79,7	80,1
boolq	RC	acc	68,3	68,9	71,4	69,4	69,0	70,1	69,1	72,7	70,6	71,7	70,4	72,1	74,3	76,8	77,8	77,7	77,9	77,7
copa	CR	acc	83,0	78,0	84,0	83,0	83,0	84,0	84,0	82,0	84,0	86,0	86,0	90,0	90,0	91,0	89,0	91,0	90,0	91,0
hellaswag	CR	acc	53,3	53,2	53,4	53,7	53,6	53,6	56,9	56,5	57,1	57,3	57,4	57,4	59,2	59,4	59,6	60,1	60,1	60,2
lambada	RC	acc	70,2	66,9	67,9	68,6	68,7	68,9	72,5	69,6	70,1	71,6	71,7	72,3	74,0	71,5	71,7	72,5	72,6	73,5
multirc	RC	acc	1,7	4,7	5,1	4,5	4,7	3,4	1,6	4,7	4,8	3,7	2,9	3,4	1,9	4,6	4,3	2,9	3,6	1,8
openbookqa	CR	acc	27,8	28,2	31,4	30,2	31,8	32,4	29,4	32,4	35,2	33,4	33,0	33,8	33,0	33,2	37,0	34,4	37,4	35,8
piqa	CR	acc	77,3	76,9	77,8	77,4	77,5	77,1	78,8	77,9	78,1	78,5	77,8	77,7	79,9	79,7	79,9	79,8	80,6	80,4
race	RC	acc	37,2	37,0	36,8	38,6	38,4	38,2	38,6	39,2	40,0	40,0	39,3	40,9	40,8	40,6	41,4	41,5	42,3	41,1
rte	NLI	acc	57,0	61,4	57,0	59,6	58,8	56,3	57,4	59,6	58,1	60,3	55,6	56,7	64,3	63,2	69,0	66,8	65,3	67,9
triviaqa	QA	acc	20,5	27,6	29,6	30,2	30,6	30,8	28,7	34,0	35,0	35,9	35,8	36,2	37,7	40,1	40,8	41,2	41,2	41,8
webqs	QA	acc	4,5	13,6	17,0	17,7	19,6	20,3	6,2	17,7	21,2	23,1	25,1	24,6	7,8	18,6	22,8	26,1	26,2	28,1
wic	CL	acc	49,8	48,6	53,9	49,2	51,1	54,5	47,5	47,8	54,4	52,5	55,6	54,9	43,3	50,6	53,0	53,1	55,0	55,5
winogrande	CR	acc	64,6	67,0	66,5	67,6	68,5	67,0	67,3	68,7	69,9	69,6	71,3	71,3	70,0	71,7	73,8	72,2	74,2	74,9
wsc	CR	acc	38,5	37,5	36,5	36,5	36,5	36,5	63,5	44,2	40,4	39,4	36,5	36,5	57,7	63,5	55,8	52,9	44,2	44,2
All Tasks Average			47,6	48,7	50,0	49,9	50,2	50,3	51,1	51,6	52,5	52,8	52,5	53,1	53,4	55,7	56,6	56,4	56,3	56,5

Example prompts for different task categories

```
# Classification example:
#   - wic with Likelihood-based accuracy metric
prompt = """
Sentence 1: Let's break for lunch.
Sentence 2: A man broken by the terrible experience of near-death.

Question: Is the word 'break' used in the same way in the two sentences above?

Answer:"""
expected_completion = " no"  # evaluate against the following completions: [yes/no]


# Closed-Book Question Answering example:
#  - triviaqa with Exact Match accuracy metric checked against a list of correct answers
prompt = """
Question: What type of leaves does a koala feed on?

Answer:"""
expected_completion = " Eucalypti"  # the following are also correct answers: [Eucalyptus/Gum trees/Gum-tree/Ευκάλυπτος]


# Commonsense Reasoning example:
#  - copa with Likelihood-based accuracy metric
prompt_1 = "The man perceived that the woman looked different because the woman got her hair cut."
prompt_2 = "The man perceived that the woman looked different because the woman wore a bracelet."  # correct answer


# Natural Language Inference example:
#  - rte with Likelihood-based accuracy metric
prompt = """
The number of Danes opposed to swapping the krone for the euro has increased slightly to 35.3 percent, up from 34.6 percent in April, according to a poll published on Thursday by Danske Bank.

Question: The introduction of the euro has been opposed. True or False?

Answer:"""
expected_completion = " True"  # evaluate against the following completions: [True/False]


# Reading Comprehension example:
#  - xquad_en with Exact Match accuracy metric
prompt = """
Between Bingen and Bonn, the Middle Rhine flows through the Rhine Gorge, a formation which was created by erosion. The rate of erosion equaled the uplift in the region, such that the river was left at about its original level while the surrounding lands raised. The gorge is quite deep and is the stretch of the river which is known for its many castles and vineyards. It is a UNESCO World Heritage Site (2002) and known as "the Romantic Rhine", with more than 40 castles and fortresses from the Middle Ages and many quaint and lovely country villages.

Question: What flows between Bingen and Bonn?

Answer:"""
expected_completion = "Middle Rhine"
```

Example prompt in few-shot study

```
# Few-shot prompting example
#  - boolq dataset with 2-shot prompt
prompt="""
Central government -- A central government is the government of a nation-state and is a characteristic of
a unitary state. This is the same thing as a federal government which may have distinct powers at various
levels authorized or delegated to it by its member states, though the adjective 'central' is sometimes
used to describe it. The structure of central governments varies. Many countries have created autonomous
regions by delegating powers from the central government to governments at a subnational level, such as a
regional, state or local level. Based on a broad definition of a basic political system, there are two or
more levels of government that exist within an established territory and govern through common
institutions with overlapping or shared powers as prescribed by a constitution or other law.
Question: is national and federal government the same thing?
Answer: yes

White House -- The White House is the official residence and workplace of the President of the United
States. It is located at 1600 Pennsylvania Avenue NW in Washington, D.C., and has been the residence of
every U.S. president since John Adams in 1800. The term White House is often used as a metonym for the
president and his advisers, as in ``The White House announced that ... ''.
Question: is the white house in the state of washington?
Answer: no

NCIS: New Orleans (season 4) -- The fourth season of NCIS: New Orleans premiered on September 26, 2017 on
CBS. The series continues to air following Bull, Tuesday at 10:00 p.m. (ET) and contained 24 episodes.
The season concluded on May 15, 2018.
Question: is ncis new orleans over for the season?
Answer:"""
```

Contact

E-Mail:	support@aleph-alpha.com	**Twitter:**	Aleph__Alpha
Website:	www.aleph-alpha.com	**Medium:**	Aleph Alpha Blog
Playground:	app.aleph-alpha.com	**YouTube:**	Aleph Alpha Channel
LinkedIn:	Aleph Alpha		