



CALO Test Plan

CALO Team

Point of Contact:

C. Raymond Perrault
Deputy-PI
AI Center, SRI International
333 Ravenswood Ave
Menlo Park, CA 94025
Email: perrault@ai.sri.com
Tel: (650) 859-6470

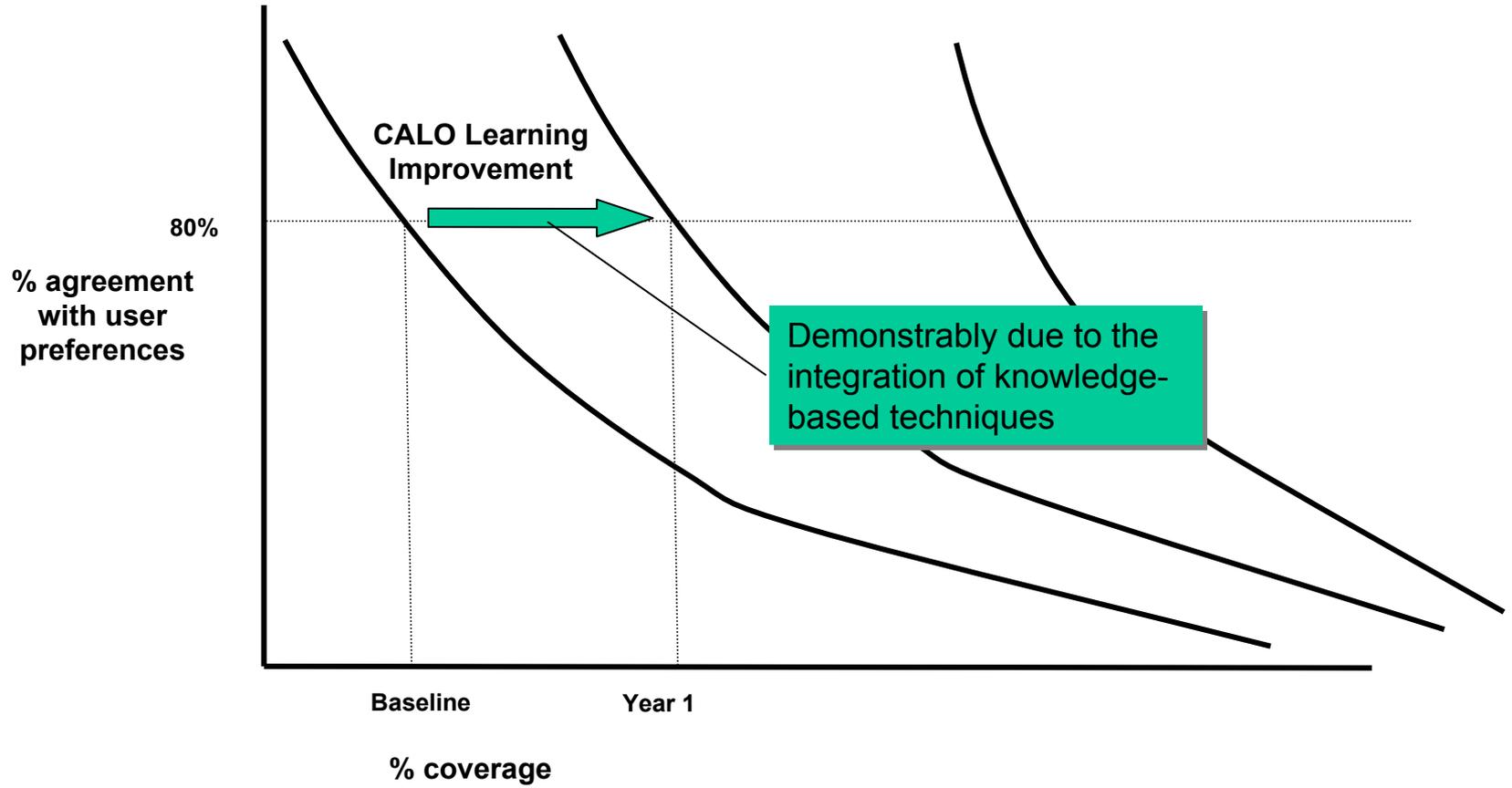
CALO Test Program

- ◆ **Although the objective of the CALO project is to develop and integrate a broad range of technologies in support of a new generation of Personal Assistants, the focus of the CALO Test Plan is on CALO's ability to learn.**
- ◆ **The test program has two phases**
 - **In Year 1, tests will be made of the application of existing and new learning techniques to the learning of preferences in CALO's application domains, but in isolation of other technologies**
 - **In Years 2 and beyond, the test will assess the learning capabilities of an end-to-end system**

Year 1 Test: Preference Learning

- ◆ **The outcome of a learning experiment is frequently represented as a curve plotting accuracy vs coverage (see next slide)**
 - **Once trained, a learning system is tested by giving it new inputs, on which it makes one of a set of possible decisions (eg which folder to file an incoming email into), or makes no decision at all.**
 - **Define**
 - **D = number of test cases on which the algorithm makes a decision**
 - **T = total number of test cases**
 - **C = number of test cases on which the algorithm makes the correct decision**
 - **Coverage = D / T**
 - **Accuracy = C / D**
 - **Generally, coverage and accuracy are inversely related**
 - **Learning algorithm A is better than algorithm B if its curve lies to the right of B's**
 - **One way to define a single representative number for a curve is as the value of coverage for a fixed accuracy. This is particularly useful if the algorithm is not useful in practice unless the accuracy is sufficiently high.**
- ◆ **The objective of the CALO Year 1 test is to demonstrate, for a number of learning methods, improvement in coverage at 80% accuracy**

Year 1: Preference Learning



Years 2-5 End-to-End Test

- ◆ **Patterned after standardized achievement tests (e.g., the Advanced Placement Exams)**
- ◆ **Each problem consists of a brief scenario followed by questions and exercises based on that scenario**
- ◆ **All questions and exercises have objectively determined “best answers”**
 - Usually there is only one best answer, but occasionally there may be several
 - Partial credit can be awarded
- ◆ **Exam is administered by an external examiner (not a member of the development team)**
- ◆ **Problems are shown on the following slides**
 - Scoring is on a 1 – 10 point scale, 10 being the most difficult
 - CALO’s exam score = sum of the points for each correct solution
- ◆ **Test protocol**
 - CALO learns over user data over several months
 - The test is administered to two versions of CALO
 - One containing the learned knowledge
 - One in which that knowledge has been ablated.
 - The test is repeated annually
- ◆ **The object of the research program is to demonstrate improvement over the years in**
 - The absolute level of performance of system with learning
 - The difference between system with and without learning

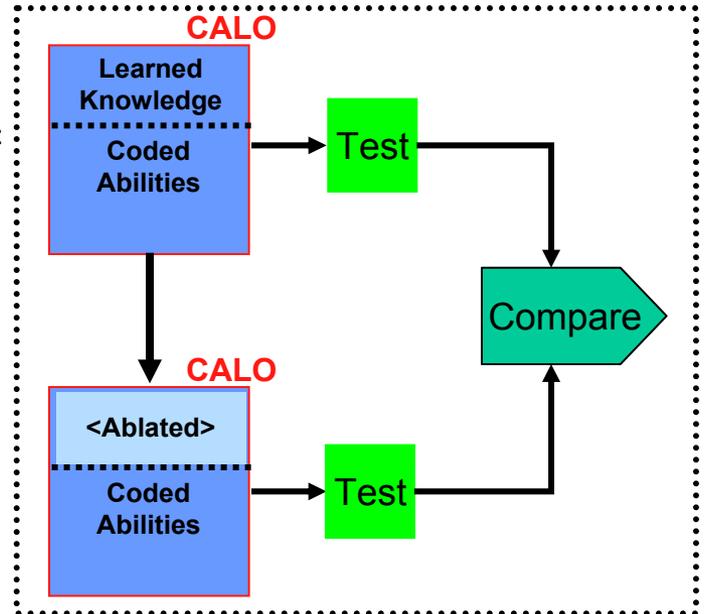
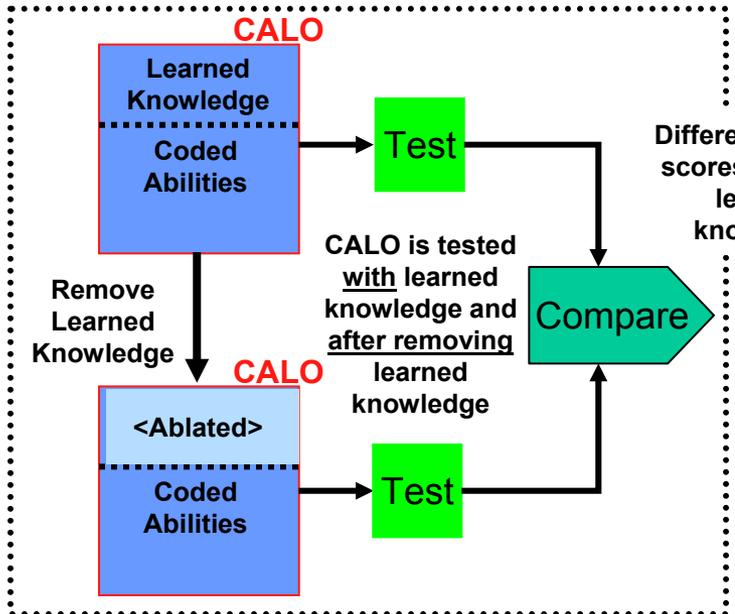
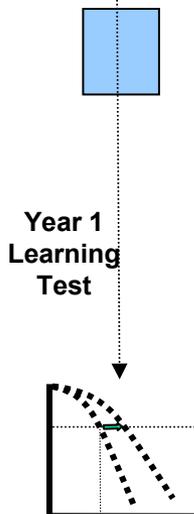
CALO Test Plan

Measure CALO performance improvement due to learning every 12 months



CALO learns by interacting with users (developers) over 24 months

CALO learns by interacting with users (developers) over 12 additional months



- CALO at 36 months incorporates new coded abilities and additional knowledge learned over the preceding 12 months
- Test is identical
- Entire process is repeated every 12 months

Project Setup Problem

Scenario A: Set up a high-priority project to deliver a program plan two weeks from today. The participants are Bob (designer), Ted (programmer), Carol (project leader) and Alice (boss). Their schedules are available to you. Set up meetings to discuss strategy, budget, specific objectives and schedule. Not everyone needs to attend all the meetings.

- A.1** Carol is now unable to attend the budget meeting as scheduled. What does CALO recommend?
- A.2** Ted is now unable to attend the budget meeting as scheduled. What does CALO recommend?
- A.3** Bob is now unable to attend the strategy meeting as scheduled. What does CALO recommend?
- A.4** Alice needs to travel to Boston during the first week of the project. How does CALO handle this?
- A.5** 15 minutes before the weekly budget meeting, Carol is still at home. What does CALO do?
- A.6** The project budget overruns by 10%. What does CALO do?

Project Execution Problem

Scenario B: We're in a project meeting. We will be discussing technical progress and some budget issues. We absolutely have to pick one of the system architecture alternatives this time. This is Bob's first meeting as a member of the team; he's a software developer.

- B.1** Which architecture was picked?
- B.2** Who wrote on the board at 2 PM?
- B.3** Send Bob the critical-path slide we edited in the meeting.
- B.4** Who had questions about slide10?
- B.5** Was John upset about his action item?
- B.6** Is the project on budget?
- B.7** In anticipation of user needs, what files does CALO put on the user's desktop?
- B.8** Does CALO identify any action items for after the meeting?

Equipment Purchase Problem

Scenario C: I need to purchase a laptop computer with a clock speed of 1.5GHz, 500MB of RAM and a 200GB hard drive. I can spend up to \$1500. Find me an appropriate machine and get approval to purchase it (e.g., retrieve 2 bids from competing vendors, complete the online purchase requisition form, get authorizations from management). I need to place the order within 4 hours.

- C.1** Did CALO find a computer that met the requirements?
- C.2** Did CALO get two bids?
- C.3** Did CALO get two authorizations?
- C.4** Did CALO complete the requisition form?
- C.5** Did CALO meet the task deadline?